ON OPTIMAL ENTROPY-CONSTRAINED SCALAR QUANTIZATION

by

András György

A thesis submitted to the
Department of Mathematics and Statistics
in conformity with the requirements for
the degree of
Master of Science (Engineering)

Queen's University
Kingston, Ontario, Canada
December, 2000

Acknowledgments

First of all I would like to thank my supervisor and friend, Dr. Tamás Linder. His expertise in source coding and constant support are certainly appreciated. He was always there to help, but he also gave me the freedom I needed to work efficiently. I am also grateful for the friendly atmosphere with which he and his family provided me during my stay in Canada.

I would also like to thank Dr. László Györfi, who started my research carrier, and who put me in touch with Dr. Linder.

Finally, I would like to acknowledge the financial support of Dr. Tamás Linder, the Department of Mathematics and Statistics, the School of Graduate Studies, the Ontario Graduate Scholarship (OGS) Program of the Ministry of Training of Ontario, and the Soros Foundation.

Abstract

Optimal scalar quantization subject to an entropy-constraint is studied.

First the problem of finding analytically an optimal entropy-constrained scalar quantizer (ECSQ) is considered. For a wide class of difference distortion measures including rth power distortions with $r > 0$, it is proved that if the source is uniformly distributed over an interval, then for any entropy constraint R (in bits), an optimal quantizer has $N = \lceil 2^R \rceil$ interval cells such that $N - 1$ cells have equal length d and one cell has length $c \leq d$. Based on this result, a parametric representation of the minimum achievable distortion $D_h(R)$ as a function of the entropy constraint R is obtained for a uniform source. Contrary to earlier expectations, the $D_h(R)$ curve turns out to be nonconvex in general. In particular, for the squared error distortion it is shown that $D_h(R)$ is a piecewise concave function.

The structural properties of optimal ECSQs for more general source distributions are also investigated. In contrast to the case of optimal fixed-rate quantization, a simple example shows that when searching for a mean-square optimal ECSQ for a source with a discrete distribution, it is generally incorrect to restrict the attention to regular quantizers, that is, to quantizers with interval cells. We present new results concerning the structure of ECSQs. One of our main results shows that for continuous sources and distortion measures of the form $d(x, y) = \rho(|x - y|)$, where ρ is a nondecreasing convex function, any finite-point ECSQ can be "regularized" so

that the resulting regular quantizer has the same entropy and equal or less distortion. This result implies that any ECSQ can be arbitrarily well approximated by finite-level regular ECSQs. Introducing the concept of almost regular quantizers (such quantizers have interval cells and they are defined almost everywhere with respect to the source distribution) we extend this result to infinite-point quantizers. We show that for any quantizer there is an almost regular quantizer with not greater entropy and equal or less distortion.

The problem of the existence of optimal ECSQs is also addressed. We prove for convex difference distortion measures and for a wide class of continuous source densities that an almost regular optimal ECSQ exists for any rate constraint $R \geq 0$. Specialized to the case of squared error distortion, we show that an optimal regular ECSQ exists for all $R \geq 0$ if the source has a piecewise monotone and piecewise continuous density. In particular, such an optimal ECSQ has a finite number of cells if the source density has a bounded support.

Contents

List of Figures

Chapter 1

Introduction

In this chapter the problem of entropy-constrained quantizer design is introduced. Section 1.1 describes the main concepts and motivation for the problem. Literature overview follows in Section 1.2. Our contribution is described in Section 1.3, and the latter section also contains a detailed overview of the thesis.

1.1 Motivation

The process of converting a stream of analog or very high-rate discrete data into a data stream with lower rate suitable for transmission or storage in digital form is called data compression or source coding. A mathematical analysis of data compression was first provided by Shannon [33] in 1948, and since then – accelerated by the demands of the rapidly growing telecommunication industry – extensive work has been done on this subject. A simple model for data compression is shown in Figure 1.1. A source generates a data sequence $\{X_i\}$, this is compressed by the source encoder $\mathcal{E}$, and then the compressed data $\mathcal{E}(X)$ is sent through a channel. Since in this thesis the emphasis is on the theoretical limitations of the source coder, we assume that the

Figure 1.1: A simple model for the source coding procedure.

channel is noiseless. This assumption is common in the source coding literature, and it makes the problem at hand more amenable to deeper analysis. At the other end of the channel, the transmitted data is decoded by the source decoder $\mathcal{D}$, and the recovered data $\widehat{X} = \mathcal{D}(\mathcal{E}(X))$ is passed to the receiver.

Data compression methods fall into two main types. The first category is lossless data compression where the reconstructed data has to match the original. However, in many situations we only require the reconstructed data to match the original one within a certain fidelity. In this case we talk about lossy data compression. In some cases, for example, if the source has a continuous output, it is even impossible to have perfectly matching reconstructed data. In this thesis we are concerned with the study of the lossy coding of real-valued sources.

Scalar (or zero-memory) quantization is the simplest method for the lossy coding of an information source with real-valued outputs. A scalar quantizer maps each output of the source to one of finitely or countably many possible quantizer output values. The quantizer outputs can be further encoded using lossless compression, and it turns out that a scalar quantizer followed by variable-rate lossless coding (entropy coding) can perform remarkably well, which makes this method popular in applications where implementation complexity is a decisive factor. For this reason, entropy coded scalar quantization is in the core of many speech and image coding algorithms, and is extensively used in mobile communication, where its simple yet efficient structure can reduce the complexity and power constraint of portable devices.

The two main quantities characterizing a scalar quantizer Q are its distortion and rate. The distortion $D(Q)$ is the average distortion between the source and the quantizer output (induced by some distortion measure). If the quantizer is followed by a fixed-rate lossless code (the outputs are represented by codewords of the same length), we talk about fixed-rate quantization, and the rate is defined as the logarithm of the number of the possible quantizer outputs. If Q is followed by entropy coding, the rate is usually defined as the entropy $H(Q)$ of the output of Q, and we talk about entropy-constrained quantization. One would like to make both $H(Q)$ and $D(Q)$ as small as possible, but these quantities are inversely related. A natural design problem is then to minimize $D(Q)$ subject to an entropy constraint $H(Q) \leq R$.

Some natural questions arise at this point. What is the lowest possible distortion, denoted by $D_h(R)$, of any scalar quantizer with output entropy less than or equal to R? Is there a quantizer which achieves this minimum distortion and also satisfies the entropy constraint? (Such a quantizer is called an optimal entropy-constrained scalar quantizer (ECSQ).) Moreover, if we know that an optimal ECSQ exists, how can we find it? What are the main structural properties of optimal ECSQs which can lead to efficient algorithms for designing good quantizers?

These questions have been extensively studied and mostly answered in fixed-rate quantization. On the other hand, not much is known about the existence and structure of optimal ECSQs. The $D_h(R)$ curve is also unknown except for the case of an exponentially distributed source and the squared error distortion. This thesis is devoted to provide some answers to these problems.

1.2 Literature overview

In this section we are concerned with those parts of the literature which are directly connected to our investigation. For a general survey on quantization the reader is referred to [13] and [20].

Fixed-rate quantization has been well studied; both practically and theoretically, a high level of understanding has been achieved. Necessary conditions (namely, the nearest neighbor and the centroid conditions) for the optimality of fixed-rate quantizers were given for the scalar case by Lloyd [24], popularized by Max [26], and generalized to higher dimensions by Linde *et al.* [22], yielding efficient algorithms for optimal quantizer design [24, 22]. Optimality of Lloyd's method was proved for scalar sources with log-concave densities by Trushkin [35] and Kieffer [18, 19]. Regarding existence, among others Pollard [30], Abaya and Wise [1], and Sabin [32] proved that an optimal N-level quantizer always exists under rather general conditions on the source distribution and the distortion measure.

On the other hand, it appears that much less is known theoretically about entropy-constrained quantization. Although it has been long known that for scalar sources entropy coded infinite-level uniform quantizers are asymptotically optimal for high rate [10, 9, 37], very few and limited results are known about the structure of optimal entropy-constrained scalar or vector quantizers at moderate and low rates, and the existence of optimal entropy-constrained vector quantizers (ECVQs) is an open question in general. An optimal ECVQ is not necessarily a nearest neighbor quantizer, and thus in general it is incorrect to restrict the attention to regular quantizers when searching for an optimal ECVQ. A nearest neighbor type condition does hold for an optimal ECVQ which achieves the lower convex hull of $D_h(R)$, implying that such a quantizer can be assumed to be regular [4], that is, for any such quantizer there is a regular quantizer with the same entropy and distortion.

In general, efforts have focused on finding necessary conditions for the optimality of a regular ECSQ with a fixed number of output points N for the squared-error distortion [3, 7, 21]. These conditions give rise to practical algorithms for designing an ECSQ with a fixed number of output points [3, 27, 7, 21, 5]. The algorithm in [5] can also be used in higher dimensions, where it finds a locally optimal ECVQ.

Sufficient conditions for the existence of an optimal quantizer among all N-point quantizers were given in [21]. However, to our knowledge, no general results corresponding to the existence of optimal ECSQs are known. Similarly, the operational distortion rate curve $D_h(R)$ is not known for most situations. The only exception is the case of an exponentially distributed source and mean squared distortion considered by Berger [3]. He derived an analytic expression for $D_h(R)$ based on the observation that the infinite-level uniform quantizer is optimal for the exponential distribution.

For the case of a uniformly distributed source, Farvardin and Modestino [7] experimentally determined the optimal quantizer using a numerical approach for the case of squared error distortion.

1.3 Contribution and thesis overview

In this thesis we address some fundamental open problems in entropy-constrained quantization.

Chapter 2 contains the necessary background information for understanding the later parts of the thesis and also contains a more detailed literature overview.

As we mentioned before, finding optimal ECSQs analytically for a given source distribution is a rather hard problem. However, a tractable but still interesting case is when the source is uniformly distributed over a finite interval. In Chapter 3 we determine analytically the optimal ECSQ for this situation. We allow a rather wide class of difference distortion measures including rth power distortions $d(x, y) = |x-y|^r$ with $r > 0$, and distortion measures of the form $d(x, y) = \rho(|x - y|)$, where ρ is a nonnegative, strictly increasing, and convex function. A parametric expression for the $D_h(R)$ curve is also given, and some of its properties are explored. Our result rigorously proves that the ECSQs found for the uniform distribution by Farvardin and Modestino [7] (using a numerical approach) are indeed optimal. The result also agrees with the one independently obtained by Topsøe [34] in a prediction context. (The material of this chapter is published in [14] and [15].)

In Chapter 4 the existence and structural properties of optimal ECSQs are analyzed. A simple example given at the beginning of the chapter shows that when searching for a mean-square optimal ECSQ for a source with a discrete distribution, it is generally incorrect to restrict the attention to quantizers with interval cells. One of our main goals in this part is to show that such anomalous behavior is not possible if the source is continuous.

In Section 4.1 we present new results concerning the structure of ECSQs. The first one shows that for continuous sources and distortion measures of the form $d(x,y) = \rho(|x-y|)$, where ρ is a nondecreasing convex function, for any finite-level ECSQ there is a regular quantizer (that is, a quantizer with interval cells and codepoints lying inside the corresponding cells) which has the same cell probabilities and equal or less distortion. This result implies that any ECSQ can be arbitrarily well approximated by finite-level regular ECSQs. Then we introduce the concept of almost regular quantizers. Such a quantizer has interval cells and it is defined almost everywhere with respect to the source distribution. We prove that any ECSQ can be replaced with an almost regular one which has equal or less entropy and distortion.

We further study almost regular quantizers in Section 4.2. It turns out that for convex difference distortion measures and for continuous source distributions, there exists an optimal ECSQ which is almost regular for any rate constraint $R \geq 0$. On the other hand, among all ECSQs with at most N codepoints, the existence of a regular optimal ECSQ is proved.

The results are specialized to the practically most interesting case of squared error distortion and sources with piecewise monotone densities in Section 4.3. We prove that in this case an optimal ECSQ which is regular always exists, and it has only finitely many codepoints in every bounded interval.

A short discussion of regularity in Section 4.4 closes this chapter.

Finally, conclusions are drawn and further work is described in Chapter 5.

Chapter 2

Quantization

2.1 Preliminaries

Definition 2.1 *A k-dimensional N-point (or N-level)* quantizer *Q is a measurable mapping of a subset $\mathcal{S}$ of the k-dimensional Euclidean space $\mathbb{R}^k$ into a finite or countably infinite set of distinct k-dimensional real vectors $\{c_1, \ldots, c_N\} \subset \mathbb{R}^k$ called the* codebook *of Q. The elements of the codebook, c_i, are called the* codepoints *of Q, and the associated sets $S_i = \{x : Q(x) = c_i\}$, $i = 1, \ldots, N$ are called the* quantization cells *of Q.*

In case $k = 1$, we talk about *s*calar quantization, and we assume that $\mathcal{S} = (\sigma, \tau)$ is a subinterval of the real line $\mathbb{R}$. Here we allow the possibility that $\sigma = -\infty$ and $\tau = \infty$. In case the codebook is not finite, we formally define $N = \infty$ and call Q an *infinite-point quantizer* (as we will see later, infinite-point quantizers only make sense as source coders if the output is followed by entropy coding).

For convenience, we will use the following convention for indexing the cells and codepoints of an infinite-point quantizer Q if a certain ordering among the codepoints can be established:

- The index set will be the set of positive integers if the codepoints of Q can be reindexed such that $c_1 < c_2 < \cdots$ or $c_1 > c_2 > \cdots$.
- All integers will be used if the set of codepoints can be *linearly ordered*, that is, if the original codepoints can be reindexed such that $\cdots < c_{-1} < c_0 < c_1 < \cdots$.

The codepoints c_i and the associated quantization cells $S_i = \{x : Q(x) = c_i\}$, $i = 1, \ldots, N$ completely characterize Q since the S_i form a partition of $\mathcal{S}$ and

$$Q(x) = c_i \quad \text{if and only if} \quad x \in S_i.$$

A more detailed way of quantizer implementation is shown in Figure 2.1, which is a specialized version of the model of Figure 1.1. The source encoder $\mathcal{E}$ and decoder $\mathcal{D}$ are divided into two parts. First $\mathcal{E}_1 : \mathcal{S} \to \{1, \ldots, N\}$ maps the quantizer input X to one of the index values. Then the generated index I is encoded by the uniquely decodable code $\mathcal{E}_2 : \{1, \ldots, N\} \to \{0,1\}^*$ to a sequence of bits $\{b_i\}$, where $\{0,1\}^*$ denotes the set of all binary sequences of finite length. Let $\mathcal{B} = \{\mathcal{E}_2(I) : I \in \{0, \ldots, N\}\}$ be the range of $\mathcal{E}_2$. After transmission through the channel, $\mathcal{D}_2 : \mathcal{B} \to \{1, \ldots, N\}$ decodes the original index value, and finally $\mathcal{D}_1 : \{1, \ldots, N\} \to \{c_1, \ldots, c_N\}$ calculates the quantizer output

$$Q(X) = \mathcal{D}_1(\mathcal{D}_2(\mathcal{E}_2(\mathcal{E}_1(X)))).$$

Now assume that $X = (X_1, \ldots, X_k)$ is a random vector taking values in $\mathcal{S}$, and let μ denote its distribution. On the one hand, we want the quantizer output $\widehat{X} = Q(X)$ (which is an approximation of X, recall the model of Figure 1.1) to be close to X. On the other hand we want the approximation $\widehat{X}$ to be sufficiently simple so that the average number of bits generated by $\mathcal{E}_2$ is small. Basically this number is regarded as the *rate* of the quantizer Q. If $\mathcal{E}_2$ encodes large blocks of indexes corresponding to successive source letters together, then the rate can be further decreased. If $\mathcal{E}_2$

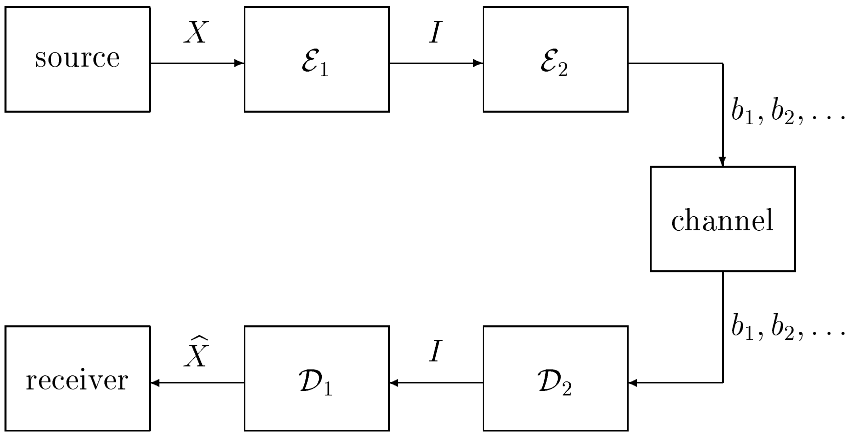

Figure 2.1: Detailed structure of a quantizer.

is a fixed-rate lossless code, we talk about *fixed-rate* quantization (discussed in Sections 2.3). If $\mathcal{E}_2$ is a variable-rate lossless code, the corresponding procedure is called *variable-rate* quantization (discussed in Section 2.4).

To quantify the fidelity of reproduction, the distortion of Q in quantizing the random variable X is defined by the expectation

$$D(Q, \mu) := E\{d(X, Q(X))\} = \int_S d(x, Q(x))\, d\mu(x)$$

where the *distortion measure* $d(\cdot, \cdot)$ is a nonnegative measurable function of two k-dimensional real vector variables. When the distribution μ is clear from the context, the short notation $D(Q)$ will be used. Under general conditions on μ and d [12, 30, 1, 32], we have $E\{d(X, c)\} < \infty$ for all $c \in \mathbb{R}^k$ and for any Borel set $S \subset \mathbb{R}^k$ there is a $c^* \in \mathbb{R}^k$ (the "generalized centroid" of S) minimizing the distortion over S, i.e.,

$$\int_S d(x, c^*) d\mu(x) = \inf_{c \in \mathbb{R}^k} \int_S d(x, c) d\mu(x) < \infty. \tag{2.1}$$

It is also shown in [12] that if the distortion measure $d(x, y)$ is strictly convex in y, then c^* is unique under some non-restrictive conditions. In case of the *mean squared error* (MSE) distortion, that is, when $d(x, y) = ||x - y||^2$, where $|| \cdot ||$ denotes the

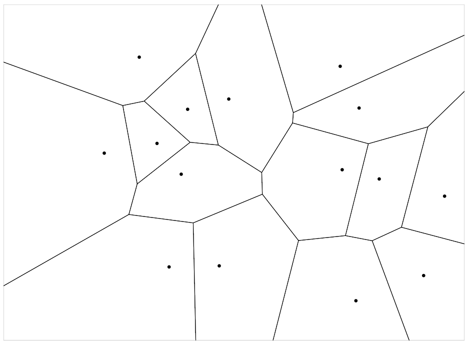

Figure 2.2: Regular vector quantizer example: $k = 2$, $N = 16$.

Euclidean norm, it suffices to assume that $E\{||X||^2\} < \infty$. Then (2.1) holds with a unique minimizing c^* for all Borel sets $S \subset \mathbb{R}^k$ with $\mu(S) > 0$.

2.2 Regular Scalar Quantizers

If Q is a finite-point quantizer whose cells S_i are convex subsets of $\mathcal{S}$, and $c_i \in S_i$ for all i, then the quantizer Q is called *regular*. Interestingly, the cells of such quantizers are always convex polytopes (see Lemma A.1 in the Appendix). An example of regular vector quantizers with convex polytope cells is shown in Figure 2.2.

Regular quantizers are of particular interest. For example, if d is a *difference distortion measure* of the form $d(x, y) = \rho(||x-y||)$ such that ρ is nondecreasing, then for any N-point quantizer Q there is a regular N-point quantizer $\widehat{Q}$ such that $D(\widehat{Q}) \leq D(Q)$. In this sense, optimal fixed-rate quantizers[1] can be assumed to be regular for nondecreasing difference distortion measures (see Section 2.3). Furthermore, the implementation of regular quantizers is quite simple in practice. An N-point regular scalar quantizer can be implemented using $N-1$ parallel comparators. For a given input point the quantizer output can also be found by a binary code tree; in this

[1] That is, quantizers whose rate is determined by the number of their codepoints.

case approximately $\log N$ sequential comparisons are necessary. In the vector case, a regular N-point quantizer can be implemented by a parallel structure of hyperplane decision functions.

Unexpected problems arise when one wants to generalize the concept of regularity to infinite-level quantizers even in the scalar case, and the definition of such quantizers is not unified in the literature. For example, a scalar quantizer with cells

$$(-\infty, 0), [1, \infty), [0, 1/2), [1/2, 3/4), [3/4, 7/8), \ldots, [1 - 2^{-n}, 1 - 2^{-(n+1)}), \ldots$$

has convex cells and so can be thought to be regular with appropriate codepoints; however there are infinitely many codepoints in any neighborhood of the point 1. To avoid such cases an additional condition is introduced to obtain a general definition of regularity, which is automatically satisfied for all regular finite-point quantizers.

Definition 2.2 *We call a quantizer Q* regular*, if it has convex cells, the codepoints lie inside the corresponding cells, and any bounded subset of $\mathbb{R}^k$ intersects only finitely many cells of Q.*

Let Q be a regular finite-point scalar quantizer with codepoints $c_1 < c_2 < \cdots < c_N$. Then the corresponding interval cells $S_1, S_2, \ldots, S_N$ satisfy

$$S_1 < S_2 < \cdots < S_N$$

where, for $A, B \subset \mathbb{R}$, $A < B$ means that $a < b$ for all $a \in A$ and $b \in B$. Therefore the interval boundaries $q_i = \sup S_i, i = 1, \ldots, N-1$ satisfy

$$\sigma = q_0 \leq c_1 \leq q_1 \leq \cdots \leq q_{N-1} \leq c_n \leq q_n = \tau$$

and

$$Q(x) = c_i \quad \text{if} \quad q_{i-1} < x < q_i.$$

The points q_i $(1 \le i \le N-1)$ are assigned to either c_i or c_{i+1}, and are called the *subdivision points* or *thresholds* of Q.

Similarly, if an infinite-point quantizer Q is regular, then its codepoints can be linearly ordered as $\cdots < c_{-1} < c_0 < c_1 < \cdots$ and there exist real numbers $\cdots \le q_{-1} \le q_0 \le q_1 \le \cdots$ such that the cells of Q are intervals with endpoints q_{i-1} and q_i, and $q_{i-1} \le c_i \le q_i$

The description of regular scalar quantizers can be further simplified for nonatomic source distributions. A random variable X has a *nonatomic distribution* μ, if $\mu(\{x\}) = 0$ for all $x \in \mathbb{R}$. (Equivalently, X has a nonatomic distribution if its distribution function $F(x) = P\{X \le x\}$ is continuous.) If ρ satisfies (2.1), it is enough to describe a quantizer by its cells, since the codepoints can be chosen optimally (but not necessarily uniquely) according to (2.1) once the cells are specified. Moreover, if X has a nonatomic distribution, then the subdivision point q_i can be mapped arbitrarily to either c_i or c_{i+1}, the distortion of Q remains the same. Hence we can assume that the cells of Q are of the form $S_i = [q_{i-1}, q_i)$, and in this case Q is determined by the subdivision point vector $(q_1, \ldots, q_{N-1})$ in the finite case (recall that $q_0 = \sigma$ and $q_N = \tau$), and by the infinite-dimensional vector $(\ldots, q_{-1}, q_0, q_1, \ldots)$ in the infinite case. We can also assume that Q does not have two interval cells of probability zero with a common endpoint (in this case we can merge these two cells without changing the entropy or the distortion). Throughout we suppose this assumption to hold.

2.3 Fixed-Rate Quantization

If the output of the quantizer is encoded with fixed length codewords, that is, $\mathcal{E}_2$ is a fixed-rate block code in the model of Figure 2.1, we call the resulting mapping a *fixed-rate* quantizer. Then to encode the N different output values, $\lceil \log N \rceil$ bits are needed, where the logarithm is base 2. Since these symbols represent k random variables, the rate of the quantizer is defined as

$$r_f = \frac{1}{k} \log N.$$

If large blocks of indexes are encoded together by $\mathcal{E}_2$, the average number of bits used for one random variable can be arbitrary close to r_f. Since the rate means the number of bits necessary to represent each element of the codebook, a fixed-rate quantizer has finite rate if and only if N, the number of codepoints, is finite.

An N-point fixed-rate quantizer is called optimal if it has the lowest possible distortion among all N-point quantizers. Formally:

Definition 2.3 *A fixed-rate quantizer Q^* is an optimal N-point quantizer if Q^* has N codepoints and*

$$D(Q^*) = \inf\{D(Q) : Q \textit{ is an } N\textit{-point quantizer}\}.$$

Optimal fixed-rate quantizers have been studied extensively. Necessary conditions for the optimality of fixed-rate quantizers were given for the scalar case by Lloyd [24], popularized by Max [26], and generalized to higher dimensions by Linde *et al.* [22]. These conditions are known as the *Lloyd-Max conditions*:

- *Nearest neighbor condition*: given the codebook $\{c_1, \ldots, c_N\}$, the optimal partition cells satisfy

$$S_i \subset \{x : d(x, c_i) \leq d(x, c_j), j = 1, \ldots, N\}.$$

- *Centroid condition*: given the partition cells $S_1, \ldots, S_N$, the optimal codepoints c_i satisfy

$$E\{d(X, c_i)|X \in S_i\} = \inf_{c \in \mathbb{R}^k} E\{d(X, c)|X \in S_i\}.$$

Note that there are optimal quantizers which do not satisfy the Lloyd-Max conditions; these quantizers can violate the nearest neighbor condition on an input set of probability zero. The Lloyd-Max conditions have some important implications. First, they immediately suggest a simple iterative algorithm for quantizer design. The algorithm starts with some initial codebook. One iteration step is the following: given the codebook, the optimal cells are determined using the nearest neighbor condition, then a new codebook is optimally generated for these new cells using the centroid condition. The iteration is then repeated with the new codebook until some appropriate stopping condition is satisfied. This algorithm is known in one-dimension as Lloyd's Method I. [24], and in higher dimensions as the generalized Lloyd algorithm or the LBG algorithm [22]. In general, the generalized Lloyd algorithm converges to a locally optimal solution only, but some results exist proving optimality for special source distributions. The uniqueness of locally optimal quantizers and hence the optimality of Lloyd's method was proved for scalar sources with log-concave densities by Trushkin [35] and Kieffer [18, 19] for convex difference distortion measures.

On the theoretical side, among others, Pollard [30], Abaya and Wise [1], and Sabin [32] proved that an optimal N-level quantizer always exists if $d(x, y)$ and μ satisfy certain regularity conditions. Specialized to squared error distortion, an optimal N-point quantizer with finite distortion exists if and only if $E\{||X||^2\} < \infty$. The proof is based on the fact that a nearest neighbor quantizer is essentially determined by its codebook $\{c_1, \ldots, c_N\}$, and thus a continuity-compactness argument can be used to show the existence of an optimal codebook.

For nondecreasing difference distortion measures, the necessary conditions for optimality imply that for any N-point quantizer Q there is a regular N-point quantizer which has less than or equal distortion. Therefore, an optimal N-point quantizer Q can be assumed to be regular.

2.4 Entropy-Constrained Quantization

Fixed-rate quantizers represent all output points with codewords of the same length. On the other hand, using *variable-rate* encoding (entropy coding) of the outputs (that is, when in the model of Figure 2.1 $\mathcal{E}_2$ is a variable-rate code), further compression can be obtained yielding an expected codelength which can be much smaller. Even more compression can be achieved by applying variable-rate encoding to large blocks of the quantizer output (in this situation $\mathcal{E}_2$ is a variable-rate block code). Therefore, if variable-rate encoding of the output of the quantizer is to be used, a new measure of the rate is required.

Let Y be a discrete random variable taking its values from a finite or countably infinite set with probabilities $p_1, p_2, \ldots$. Then the entropy of Y is defined as

$$H(Y) = -\sum_i p_i \log p_i.$$

Now assume that $\mathcal{Y} = Y_1, Y_2, \ldots$ is a discrete stationary source. Then by the *Source Coding Theorem* (see, e.g., [8]), the expected per letter codelength L of any uniquely decodable lossless binary code for the random vector $(Y_1, \ldots, Y_n)$ satisfies

$$\frac{1}{n} H(Y_1, \ldots, Y_n) \leq L$$

where $H(Y_1, \ldots, Y_n)$ is the entropy of the vector random variable $(Y_1, \ldots, Y_n)$. If the block size $n \to \infty$, then $\frac{1}{n} H(Y_1, \ldots, Y_n)$ decreases to the source entropy

$$H(\mathcal{Y}) = \lim_{n \to \infty} \frac{1}{n} H(Y_1, \ldots, Y_n).$$

Moreover, there are uniquely decodable prefix codes (for example the Shannon-Fano code [6]) for which the expected per letter codelength L satisfies

$$\frac{1}{n} H(Y_1, \ldots, Y_n) \leq L < \frac{1}{n} H(Y_1, \ldots, Y_n) + \frac{1}{n}. \tag{2.2}$$

This shows that encoding large blocks of $\mathcal{Y}$ together, the minimum expected per letter codelength $H(\mathcal{Y})$ can be arbitrarily approached. If $\mathcal{Y}$ is a sequence of independent and identically distributed random variables, then $H(\mathcal{Y}) = H(Y_1)$.

Therefore, assuming that the input sequence of the quantizer is independent and identically distributed, with a generic random variable X, an appropriate measure of the rate of the quantizer is $\frac{1}{k}H(Q)$, where

$$H(Q) = H(Q(X)) = -\sum_{i=1}^{N} P\{X \in S_i\} \log P\{X \in S_i\}.$$

This quantity is called the *entropy-constrained rate* of Q in quantizing the k-dimensional random variable $X = (X_1, \ldots, X_k)$. Since the logarithm is of base 2, $H(Q)$ is measured in bits.

Definition 2.4 *A vector quantizer whose rate is measured by* $\frac{1}{k}H(Q)$ *is called an* entropy-constrained vector quantizer *(ECVQ).*

For any $R \geq 0$ let $D_h(R)$ denote the lowest possible distortion of any quantizer with output entropy not greater than R. This function is formally defined by

$$D_h(R) = \inf\{D(Q) : H(Q) \leq R\}$$

where the infimum is taken over all finite or infinite-level scalar quantizers whose entropy is less than or equal to R.

Definition 2.5 *Any Q that achieves $D_h(R)$ in the sense that $H(Q) \leq R$ and $D(Q) = D_h(R)$ is called an* optimal *ECVQ.*

In situations when implementation complexity is a decisive factor, scalar (or zero memory) quantization can be used.

Definition 2.6 *A scalar quantizer whose rate is measured by $H(Q)$ is called an* entropy-constrained scalar quantizer *(ECSQ).*

In contrast to fixed-rate quantization, there is little known about optimal entropy-constrained quantization. On the one hand, it is easy to see that the centroid condition still holds for optimal entropy-constrained quantizers, since redefining the codepoints as the centroids of the cells can only decrease the distortion, while the entropy remains the same. On the other hand, the nearest neighbor condition clearly fails here, since redefining the cells as the nearest neighbor regions of the codepoints changes the cell probabilities, and thus can result in increased entropy.

Although it has been long known that entropy coded infinite-level uniform quantizers are asymptotically optimal for high rate [10] (a rigorous proof of this fact was given recently in [23]), very few and limited results are known about the structure of optimal entropy-constrained scalar or vector quantizers for moderate to low rates. As well, the existence of optimal ECVQs is an open question in general. Unfortunately, an optimal ECVQ is not necessarily a nearest neighbor quantizer, and thus in general it is incorrect to restrict the attention to regular quantizers when searching for an optimal ECVQ. Indeed, it is not hard to construct a discrete source with three real-valued outputs for which the unique optimal ECSQ is not regular at certain rates (see Example 4.1).

In general, efforts have focused on finding necessary conditions for the optimality of a *regular* ECSQ with a *fixed* number of output points N for the squared-error distortion [3, 7, 21]. These conditions give rise to practical algorithms for designing an ECSQ with a fixed number of output points [3, 27, 7, 21, 5]. The algorithm of [5] can also be used in higher dimensions, where it finds a locally optimal ECVQ. In [21] a constructive proof of Berger's [3] necessary conditions for optimality is given. Kieffer *et al.* [21] provided sufficient conditions for the existence of optimal regular scalar quantizers among all N-point quantizers for log-concave source densities. They showed that if the source has a strictly log-concave density, then this optimal quantizer is unique. Conditions were also given under which such an optimal quantizer does not exist. In [3] an equivalence between entropy-constrained scalar quantization and permutation codes is also given.

To determine the overall optimal ECSQ and the corresponding operational distortion-rate curve $D_h(R)$, one must find the optimum performance over all N. Unfortunately, this step is rather hard, even for the most common continuous source distributions. A notable exception is the case of an exponentially distributed source and the mean squared distortion considered by Berger [3]. He derived an analytic expression for $D_h(R)$ based on the observation that for the exponential distribution, the necessary conditions for optimality at any positive rate are satisfied by an infinite-level uniform quantizer. To our knowledge, this is the only case where a correct[2] analytic formula for $D_h(R)$ is known; the resulting quantizer is an infinite-point uniform quantizer.

The properties of $D_h(R)$ are in general not well understood. It is easy to see that $D_h(R)$ is a decreasing function. One can now ask whether (in analogy to the rate-distortion function) $D_h(R)$ is convex. The convexity of $D_h(R)$ is of particular interest because the lower convex hull of $D_h(R)$ plays a special role in variable-rate lossy coding. For example, the lower convex hull of $D_h(R)$ is the minimum achievable distortion in causal lossy coding of a memoryless source [28] (a source code is causal if the reproduction created by the code of the present source letter depends only on present and past source letters but not on future ones).

Also, Lloyd-Max type necessary conditions of optimality are known only for an

[2]Although a complete proof that infinite-level uniform quantizers are indeed optimal is missing, the result is widely believed to be correct.

optimal ECSQ which achieves the lower convex hull of $D_h(R)$ [4]. Assume that a quantizer Q achieves the lower convex hull of $D_h(R)$, and let λ be the negative of the right derivative of the lower convex hull of $D_h(R)$ at the point $H(Q)$. Then the quantizer Q satisfies the *modified nearest neighbor condition* with parameter λ: for μ-almost all $x \in \mathbb{R}^k$,

$$d(x, c_{\alpha(x)}) - \lambda \log P\{X \in S_{\alpha(x)}\} = \min_i d(x, c_i) - \lambda \log P\{X \in S_i\} \qquad (2.3)$$

where $\alpha(x) = i$ if and only if $Q(x) = c_i$. The proof of this interesting result is reproduced in Lemma A.2 in the Appendix.

Now for a discrete source, $D_h(R)$ is never convex since it is decreasing and piecewise constant. On the other hand, it can be shown (using the analytical expression of Berger [3]) that for an exponentially distributed source and the squared error distortion, $D_h(R)$ is convex. It has also been conjectured [28] that $D_h(R)$ is convex for a wide variety of source distributions and distortion measures. Our results for the uniform source demonstrate that $D_h(R)$ can be nonconvex even for "nice" continuous source distributions.

Chapter 3

Optimal ECSQ for a uniform source

In this chapter we determine analytically the optimal ECSQ for a source which is uniformly distributed over a bounded interval. To our knowledge this is the first result demonstrating a provably optimal ECSQ for a source with a density. The main results are stated and proved in Section 3.1, the technical proofs of some auxiliary lemmas can be found in Section 3.2.

We allow a rather wide class of difference distortion measures including rth power distortions $d(x, y) = |x-y|^r$ with $r > 0$, and distortion measures of the form $d(x, y) = \rho(|x-y|)$, where ρ is a nonnegative, strictly increasing, and convex function. The main result of the chapter is stated in Theorem 3.1; it proves that an optimal ECSQ for any rate $R \geq 0$ (measured in bits) is an $N = \lceil 2^R \rceil$-point quantizer, where $\lceil x \rceil$ denotes the smallest integer not less than x. This quantizer has $N-1$ cells of equal length d and one cell of length $c \leq d$, where d and c are uniquely determined by the requirement that the entropy constraint is satisfied with equality (the optimal quantizer is uniform if $2^R = N$, as expected). Specialized to the squared error distortion, our result rigorously proves that the ECSQs found for the uniform distribution by Farvardin and Modestino [7] (using a numerical approach) are indeed optimal. In case of the absolute-error distortion our result also agrees with a result independently obtained by Topsøe [34] in a prediction context.

Based on the analytic description of an optimal ECSQ, in Corollary 3.1 we obtain a parametric expression for $D_h(R)$. The properties of the $D_h(R)$ curve are investigated in Corollary 3.2. In general $D_h(R)$ is piecewise smooth (differentiable everywhere except at the points $R = \log N$). For the squared error distortion (and more generally for rth power distortions with $0 < r \leq 3$) we prove that $D_h(R)$ is concave in each interval $[\log(N-1), \log N]$, where $N \geq 2$ is an integer. Thus for such distortion measures, $D_h(R)$ is not convex over any interval. It also follows that in this case an optimal ECSQ achieving the lower convex hull of $D_h(R)$ exists only at rates $R = \log N$, where N is a positive integer. Moreover, since the lower convex hull of $D_h(R)$ is the minimum achievable distortion in causal source coding [28], for such distortion measures and for rates $\log(N-1) < R < \log N$, optimal causal compression of a uniform source can be achieved only via timesharing between an $(N-1)$-point and an N-point uniform quantizer.

3.1 Main results

Let X have a uniform distribution on the unit interval $(0, 1)$, and assume that the distortion measure is in the form of $d(x, y) = \rho(|x - y|)$, where $\rho : [0, \infty) \to [0, \infty)$ is strictly increasing, continuous, and $\rho(e^t)$, $t \in \mathbb{R}$, is strictly convex. Without loss of generality we also assume that $\rho(0) = 0$. Examples for such distortion measures include rth power distortions $d(x, y) = |x - y|^r$ with $r > 0$ (in this case $\rho(e^t) = e^{rt}$), and distortion measures $d(x, y) = \rho(|x - y|)$, where ρ is strictly increasing and convex on $[0,\infty)$. Another distortion measure which does not fall into either of these categories but satisfies the requirements is $d(x,y) = \log(1+|x-y|)$.

Let Q be any finite or infinite-point quantizer with cells $\{S_1,\dots,S_N\}$ and codepoints $\{c_1,\dots,c_N\}$, and define $p_i = \lambda(S_i \cap (0,1))$ for $i=1,\dots,N$, where λ denotes the Lebesgue measure. Then we can define a new quantizer $\widehat{Q}$ over $(0,1)$ which has N interval cells of length p_i, $i=1,\dots,N$, and N codepoints which are located at the midpoints of these cells (the definition of $\widehat{Q}$ outside $(0,1)$ is immaterial). Since the codepoints are placed at the midpoints of the corresponding cells, the distortion of $\widehat{Q}$ is given by

$$D(\widehat{Q}) = \sum_{i=1}^{N} \Phi(p_i) \tag{3.1}$$

where $\Phi(p)$ is defined for all $p \geq 0$ by

$$\Phi(p) = 2\int_0^{p/2} \rho(x)\,dx. \tag{3.2}$$

Since ρ is increasing, it is easy to see that for all $i=1,\dots,N$,

$$\int_{S_i\cap(0,1)} \rho(|x-c_i|)\,dx \geq \int_{-p_i/2}^{p_i/2} \rho(|x|)\,dx$$

and so $D(\widehat{Q}) \leq D(Q)$. On the other hand,

$$H(\widehat{Q}) = -\sum_{i=1}^{N} p_i \log p_i \tag{3.3}$$

so that $H(\widehat{Q}) = H(Q)$. Consequently, when searching for an optimal ECSQ for the uniform distribution over $(0,1)$, it suffices to consider interval partitions of $(0,1)$ and the associated regular quantizers with codepoints at the midpoints of the intervals. All quantizers in the rest of this chapter will be assumed to be of this type. The distortion and rate of any such quantizer are uniquely determined by the cell lengths $\{p_i;\ i=1,\dots,N\}$ through equations (3.1) and (3.3). Note that if N is finite and

$p_i = 1/N$, $i = 1, \ldots, N$, the resulting quantizer is the N-level uniform quantizer over $(0, 1)$.

If $R = 0$, the optimal ECSQ for any source distribution has only one codepoint. The next result shows that if the source has a uniform distribution, then for any rate $R > 0$ there exists an optimal finite-point ECSQ with a very simple structure.

Theorem 3.1 *Let the source X have uniform distribution over $(0,1)$ and assume that $d(x,y) = \rho(|x-y|)$, where $\rho : [0,\infty) \to [0,\infty)$ is a strictly increasing continuous function such that $\rho(e^t)$ is strictly convex. Then Q is an optimal ECSQ for a rate constraint $R > 0$ if and only if Q has $N = \lceil 2^R \rceil$ cells; one cell of length c and $N-1$ cells of length $\frac{1-c}{N-1}$, where c is the unique solution of the equation*

$$-c \log c - (1-c) \log \left(\frac{1-c}{N-1} \right) = R$$

in the interval $(0, \frac{1}{N}]$.

Remarks. Theorem 3.1 implies the intuitive result that if $R = \log N$, then the unique optimal ECSQ for the uniform source is the N-level uniform quantizer. If $R < \log N$, then $c < \frac{1-c}{N-1}$, and the optimal ECSQ is no longer unique; there are exactly N such quantizers (with interval cells). Farvardin and Modestino [7] reached the same conclusion for squared error distortion using numerical methods.

Theorem 3.1 remains valid (after rescaling) if X is uniformly distributed over an arbitrary interval (a,b). To see this, let $\widehat{X} = (X-a)/(b-a)$ and $\widehat{\rho}(x) = \rho(x/(b-a))$. Then $\widehat{X}$ and $\widehat{\rho}$ satisfy the conditions of the theorem, and a quantizer Q is optimal for X and ρ if and only if $\widehat{Q}(x) = \big[Q\big(x(b-a)+a\big) - a\big]/(b-a)$ is optimal for $\widehat{X}$ and $\widehat{\rho}$.

The proof of the theorem, given below, has two main parts. First, similar to [3, 7], the usual Kuhn-Tucker conditions of constrained optimization are used to identify necessary conditions for the optimality of an n-point ECSQ for a fixed positive integer

n. After eliminating all quantizers not satisfying these conditions, we are left with the family of n-point quantizers over $(0,1)$ which satisfy the entropy constraint with equality and whose cell lengths can take only two distinct values (these quantizers were also identified in [7]). The second, harder part of the proof consists of identifying, for a fixed n, the quantizers which have minimal distortion in this family, and then finding the optimal choice of n.

Proof of Theorem 3.1 Without loss of generality we will assume that $\rho(0) = 0$ (otherwise we can replace $\rho(x)$ by $\widehat{\rho}(x) = \rho(x) - \rho(0)$). Let Ψ be the Gish-Pierce function [10, 11] defined by

$$\Psi(p) = \begin{cases} \dfrac{\Phi(p)}{p} & \text{if } p > 0 \\ 0 & \text{if } p = 0 \end{cases}$$

where $\Phi(p) = 2\int_0^{p/2} \rho(x)\,dx$. Notice that $\Psi(p) = E\{\rho(pY)\}$ for all $p \geq 0$, where Y is a random variable that is uniformly distributed over the interval $(0, 1/2)$. Then the strict convexity of $\rho(e^t)$ implies that for all $t_1, t_2 \in \mathbb{R}$ such that $t_1 \neq t_2$, and $0 < \alpha < 1$,

$$\begin{aligned} \Psi\big(e^{\alpha t_1 + (1-\alpha)t_2}\big) &= E\big[\rho\big(e^{\alpha t_1 + (1-\alpha)t_2}Y\big)\big] \\ &= E\big[\rho\big(e^{\alpha(t_1 + \ln Y) + (1-\alpha)(t_2 + \ln Y)}\big)\big] \\ &< E\big[\alpha\rho\big(e^{t_1 + \ln Y}\big) + (1-\alpha)\rho\big(e^{t_2 + \ln Y}\big)\big] \\ &= \alpha\Psi(e^{t_1}) + (1-\alpha)\Psi(e^{t_2}) \end{aligned}$$

where ln denotes the natural logarithm. Thus $\Psi(e^t)$ is strictly convex. Since $\Psi(0) = 0$ and $\Psi(p) > 0$ for $p > 0$, the convexity of $\Psi(e^t)$ implies that $\Psi(p)$ is strictly increasing.

By the discussion preceding Theorem 3.1, we need to find nonnegative cell lengths $\{p_i;\ i = 1, 2, \ldots\}$ satisfying $\sum_i p_i = 1$, which minimize $\sum_i \Phi(p_i)$ subject to $-\sum_i p_i \log p_i \leq$

R. For all $\{p_i\}$ satisfying this entropy constraint we have

$$\begin{aligned}\sum_i \Phi(p_i) &= \sum_{i:\, p_i>0} p_i \Psi(e^{\ln 2 \log p_i}) \\ &\geq \Psi\left(e^{\ln 2 \sum_i p_i \log p_i}\right) &(3.4)\\ &\geq \Psi\left(2^{-R}\right) &(3.5)\end{aligned}$$

where (3.4) follows from Jensen's inequality (see, e.g., [2]) and the convexity of $\Psi(e^t)$, and (3.5) follows since Ψ is increasing. Thus $\Psi(2^{-R})$ is a lower bound on the distortion of any quantizer with entropy not greater than R, that is, $D_h(R) \geq \Psi(2^{-R})$. Since $\Psi(e^t)$ is strictly convex and strictly increasing, $\sum_i \Phi(p_i) = \Psi\left(2^{-R}\right)$ if and only if all positive p_i's are equal and $-\sum_i p_i \log p_i = R$. Equivalently, $R = \log N$ for some positive integer N and $p_i = 1/N$ for (say) $1 \leq i \leq N$, and $p_i = 0$ for $i > N$. The resulting quantizer is the N-level uniform quantizer over $(0,1)$ with entropy $R = \log N$ and distortion $\Psi(1/N) = D_h(\log N)$. This proves the theorem for rates $R = \log N$, where N is a positive integer.

Now consider the case $\log(N-1) < R < \log N$, where $N = \lceil 2^R \rceil$. First observe that in the infimum defining $D_h(R)$ it is enough to consider finite-point quantizers; i.e.,

$$D_h(R) = \inf\{D(Q) : Q \text{ is finite-point},\ H(Q) \leq R\}. \tag{3.6}$$

(Let Q be any infinite-point quantizer over $(0,1)$ with cell lengths $\{p_i;\, i = 1, 2, \dots\}$ and for a positive integer n, let $\widehat{Q}$ have cell lengths $\left\{p_1, \dots, p_{n-1}, \sum_{i\geq n} p_i\right\}$. Then $H(\widehat{Q}) \leq H(Q)$ and $D(\widehat{Q}) - D(Q) \leq \Phi\left(\sum_{i\geq n} p_i\right)$, and now (3.6) follows since $\sum_{i\geq n} p_i \to 0$ as $n \to \infty$.)

For a positive integer n, let $\mathcal{P}_n$ denote the n-dimensional probability simplex

$$\mathcal{P}_n = \left\{(p_1, \dots, p_n) : p_i \geq 0, i = 1, \dots, n;\ \sum_{i=1}^n p_i = 1\right\}$$

and for all $\mathbf{p} = (p_1, \ldots, p_n) \in \mathcal{P}_n$ define

$$h_n(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log p_i, \qquad \delta_n(\mathbf{p}) = \sum_{i=1}^{n} \Phi(p_i).$$

Let $R > 0$ be fixed. Since h_n and δ_n are continuous and $\mathcal{P}_n$ is compact,

$$B_{n,R} = \{\mathbf{p} \in \mathcal{P}_n : h_n(\mathbf{p}) \leq R\}$$

is compact and δ_n achieves its minimum in $B_{n,R}$. Then (3.6) implies that $D_h(R)$ is given by

$$D_h(R) = \inf_{n \geq 1} \min\{\delta_n(\mathbf{p}) : \mathbf{p} \in B_{n,R}\}. \tag{3.7}$$

Note that $\min\{\delta_n(\mathbf{p}) : \mathbf{p} \in B_{n,R}\}$ is nonincreasing in n and an optimal entropy-constrained quantizer with a finite codebook exists if and only if the infimum is achieved in (3.7) for finite n. If $\mathbf{p}^* \in \mathcal{P}_n$ minimizes δ_n over $B_{n,R}$ and it has k nonzero components, then by dropping the zero components we obtain $\mathbf{p}^{**} \in \mathcal{P}_k^+ = \{\mathbf{p} \in \mathcal{P}_k : p_i > 0, i = 1, \ldots, k\}$ which minimizes δ_k over $B_{k,R}$. Since the quantizers associated with $\mathbf{p}^*$ and $\mathbf{p}^{**}$ are identical, we can conclude that it suffices to find the positive solutions $\mathbf{p} \in \mathcal{P}_n^+$ of the constrained minimization problem

$$\begin{aligned} &\text{minimize} \quad \delta_n(\mathbf{p}) \\ &\text{subject to} \quad h_n(\mathbf{p}) \leq R, \quad \textstyle\sum_{i=1}^n p_i = 1 \end{aligned} \tag{3.8}$$

for all $n \geq 1$ such that a solution exists. Since h_n and δ_n are continuously differentiable on $B_{n,R} \cap \mathcal{P}_n^+$, we can use the Kuhn-Tucker conditions (see, e.g., [25]) to find all local minimum points in (3.8) for all n. The collection of these solutions will correspond to a simple family of finite-point quantizers where it will be possible to identify the global optimum.

By the Kuhn-Tucker conditions, if a local minimum $\mathbf{p}$ is a regular point of the constraints in (3.8) in the sense that the gradients of $h_n(\mathbf{p})$ and $\sum_{i=1}^n p_i$ are linearly independent, then there exist Lagrange multipliers $\mu \geq 0$ and $\lambda \in \mathbb{R}$ such that

$$J(\mathbf{p}, \mu, \lambda) = \delta_n(\mathbf{p}) + \mu h_n(\mathbf{p}) + \lambda \sum_{i=1}^{n} p_i$$

satisfies $\frac{\partial J}{\partial p_i} = 0$ for $i = 1, \ldots, n$. The gradients of $h_n(\mathbf{p})$ and $\sum_{i=1}^{n} p_i$ are linearly dependent if and only if $p_i = 1/n$ for $i = 1, \ldots, n$. Otherwise, the Kuhn-Tucker conditions give

$$\frac{\partial J(\mathbf{p}, \mu, \lambda)}{\partial p_i} = \rho(p_i/2) - \mu(\log e + \log p_i) + \lambda = 0, \qquad i = 1, \ldots, n$$

that is, a minimizing $\mathbf{p}$ must solve

$$\lambda = -\rho(p_i/2) + \mu(\log e + \log p_i), \qquad i = 1, \ldots, n. \tag{3.9}$$

Let λ and μ be fixed and for $p > 0$ define $v(p) = -\rho(p/2) + \mu(\log e + \log p)$. Then

$$v(2e^t) = -\rho(e^t) + \mu t \log e + \mu(1 + \log e)$$

is a strictly concave function of t (since $\rho(e^t)$ is strictly convex). Thus the equation $v(2e^t) = \lambda$ has at most two distinct solutions in t, and consequently (3.9) has at most two distinct solutions in p_i.

Hence the components of any optimal $\mathbf{p} \in \mathcal{P}_n^+$ take at most two distinct positive values. To uniquely describe such a $\mathbf{p}$ (up to permutations of the components), let c denote the smaller of the two values, let l denote the number of components equal to c, and let us specify that $c \in (0, 1/n)$ and $0 \leq l < n$. Then there are $n - l$ components equal to $d = \frac{1-cl}{n-l}$ (note that $l = 0$ if and only if all components are equal). An associated quantizer has l cells of length c and $n - l$ cells of length d, and its distortion and entropy are given respectively by

$$\delta(c, l, n) = l\Phi(c) + (n - l)\Phi(d)$$

and

$$h(c, l, n) = -lc \log c - (1 - lc) \log d.$$

Therefore, if there exist c, l, and n minimizing $\delta(c, l, n)$ subject to $h(c, l, n) \le R$, the corresponding quantizer is optimal (i.e, it achieves $D_h(R)$).

Fix $R > 0$ and assume that $\log n < R$. Since the uniform n-point quantizer has minimum distortion (namely $n\Phi(1/n)$) and maximum entropy (namely $\log n$) among all n-point quantizers, it is the optimal n-point quantizer with entropy constraint R. Since $\log n < R$, one can easily construct an $(n+1)$-point quantizer Q with n equal-length cells and one sufficiently small cell which has entropy $H(Q) < R$ and distortion $D(Q) < n\Phi(1/n)$. Hence no n-point quantizer can achieve $D_h(R)$ if $\log n < R$. Therefore we will assume that $\log n > R$ (the case $\log n = R$ was dealt with previously). We will also assume that $l > 0$, since $l = 0$ results in the n-level uniform quantizer with rate $\log n > R$.

Note that

$$\frac{\partial h(c, l, n)}{\partial c} = l(\log d - \log c) > 0 \tag{3.10}$$

since $c < d = \frac{1-cl}{n-l}$, and thus $h(c, l, d)$ is a strictly increasing function of $c \in (0, 1/n)$ for fixed l. Therefore the constraint $h(c, l, n) \le R$ can be satisfied with $c > 0$ if and only if $l \ge l_{\min}$, where the integer $l_{\min}$ is defined by

$$l_{\min} = \min\{l \ge 0 : \log(n - l) < R\}.$$

Note that $l_{\min} \ge 1$ since $\log n > R$ by assumption. Next we observe that $\delta(c, l, n)$ is strictly decreasing in c since

$$\frac{\partial \delta(c, l, n)}{\partial c} = l\left(\rho\left(\frac{c}{2}\right) - \rho\left(\frac{d}{2}\right)\right) < 0$$

(recall that ρ is strictly increasing). Thus for fixed n (such that $\log n > R$) and $l \geq l_{\text{min}}$, the unique c minimizing $\delta(c, l, n)$ subject to $h(c, l, n) \leq R$ is the unique solution of the equation

$$-lc \log c - (1 - lc) \log d = R$$

in the interval $(0, 1/n)$. Let $c(l, n, R)$ denote this solution, and denote the corresponding distortion by

$$\delta^*(l, n, R) = \delta(c(l, n, R), l, n). \tag{3.11}$$

Lemma 3.1 in the next section shows that $\delta^*(l, n, R)$ is strictly increasing in l for fixed n and R, and therefore $\delta^*(l, n, R)$ is uniquely minimized in l by the choice $l = l_{\text{min}}$, the smallest possible value of l.

Now if $\log(N-1) < R < \log N$, then $\log n > R$ if and only if $n \geq N$. For any such n, we have $l_{\text{min}} = n - N + 1$, and now $\delta^*(l, n, R)$ is minimized in l by $l = n - N + 1$. The corresponding distortion is

$$\delta^*(n - N + 1, n, R) = (n - N + 1)\Phi(c) + (N - 1)\Phi(d) \tag{3.12}$$

where c is the unique solution of the equation

$$-(n - N + 1)c \log c - (1 - (n - N + 1)c) \log d = R \tag{3.13}$$

in $(0, 1/n)$ where $d = \frac{1-(n-N+1)c}{N-1}$. Lemma 3.2 in the next section shows that $\delta^*(n - N + 1, n, R)$ is strictly increasing in n for fixed N and R, and thus it is minimized by $n = N$, the smallest possible choice for n. We can conclude that any quantizer over $(0, 1)$ with one cell of length c and $N - 1$ cells of length $d = \frac{1-c}{N-1}$, where c satisfies $-c \log c - (1 - c) \log(\frac{1-c}{N-1}) = R$, is optimal; i.e., it achieves $D_h(R)$. It also follows that any other quantizer with a different set of cell lengths is strictly suboptimal. $\square$

Using equations (3.1) and (3.3), the distortion and entropy of an N-point quantizer Q with cell lengths $p_1 = c$ and $p_i = \frac{1-c}{N-1}$, $i = 2, \ldots, N$ are given respectively by

$$D(Q) = \Phi(c) + (N-1)\Phi\left(\frac{1-c}{N-1}\right).$$

and

$$H(Q) = -c\log c - (1-c)\log\left(\frac{1-c}{N-1}\right)$$

It is easy to see that $H(Q)$ is a strictly increasing function of $c \in [0, \frac{1}{N}]$. Also, $H(Q) = \log(N-1)$ if $c = 0$, $H(Q) = \log N$ if $c = \frac{1}{N}$, and the corresponding quantizers are the $(N-1)$-level and N-level uniform quantizers, respectively. Thus Theorem 3.1 yields the following parametric description of $D_h(R)$ for the uniform distribution.

Corollary 3.1 *For X and $d(x,y)$ as in Theorem 3.1, and for any positive integer $N \geq 2$, the parametric representation of $D_h(R)$ in the interval $[\log(N-1), \log N]$ is given by*

$$R_c = -c\log c - (1-c)\log\left(\frac{1-c}{N-1}\right)$$

$$D_h(R_c) = \Phi(c) + (N-1)\Phi\left(\frac{1-c}{N-1}\right)$$

where $c \in [0, \frac{1}{N}]$.

From the parametric representation we can deduce some important properties of $D_h(R)$. For example, it immediately follows that $D_h(R)$ is everywhere continuous. Moreover, plotting $D_h(R)$ for the squared error distortion $d(x,y) = (x-y)^2$ (see Figure 3.1) suggests that the graph of $D_h(R)$ consists of smooth, concave pieces joined in a nonsmooth manner at rates $R = \log N$. The next result proves these properties of $D_h(R)$ under more general conditions.

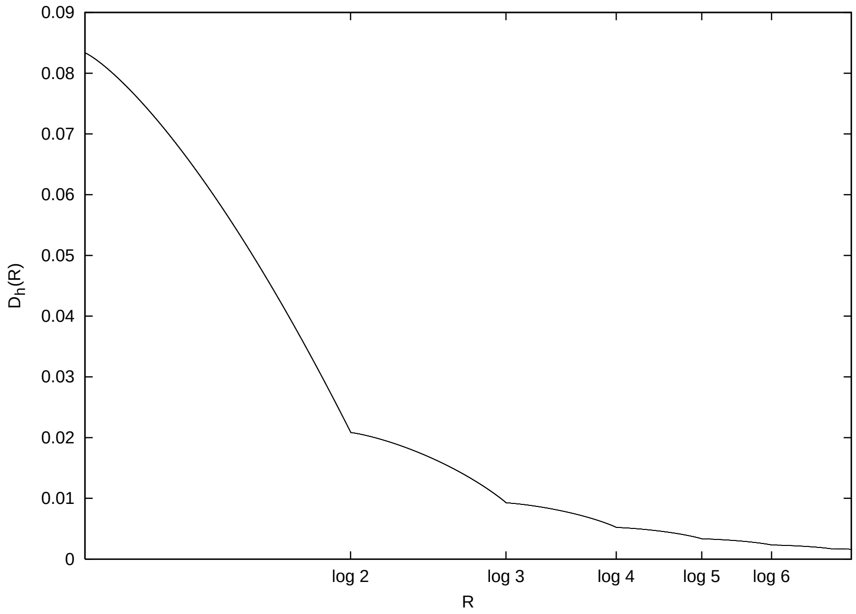

Figure 3.1: $D_h(R)$ for the uniform source and squared error distortion.

Corollary 3.2 *With the conditions of Theorem 3.1, $D_h(R)$ has the following properties:*

(i) *$D_h(R)$ is continuously differentiable on each open interval* $(\log(N-1), \log N)$, *where $N \geq 2$ is a positive integer. At $R = \log N$, the right derivative of $D_h(R)$ is zero for all $N \geq 1$, and the left derivative of $D_h(R)$ is negative for all $N \geq 2$. Thus $D_h(R)$ is not differentiable at the points $R = \log N$ for $N \geq 2$.*

(ii) *Let $d(x,y) = |x-y|^r$ be the rth power distortion with $0 < r \leq 3$. Then $D_h(R)$ is strictly concave on each interval* $[\log(N-1), \log N]$ *for $N \geq 2$.*

Remarks. The proof below also shows that (ii) cannot be improved in the sense that if $d(x,y) = |x-y|^r$ with $r > 3$, then $D_h(R)$ is no longer concave on $[0,1]$.

Part (i) of the preceding corollary implies that $D_h(R)$ is not convex. Moreover, part (ii) shows that for the squared error distortion an ECSQ achieving the lower convex hull of $D_h(R)$ exists only at the discrete rate values $R = \log N$. This fact suggests that an ECSQ which achieves the lower convex hull of $D_h(R)$ is the exception rather than the rule.

Proof of Corollary 3.2

(i) Recall that by Corollary 3.1, in the interval $[\log(N-1), \log N]$ the parametric equations of $D_h(R)$ are

$$R_c = -c \log c - (1-c) \log d$$

$$D_h(R_c) = \Phi(c) + (N-1)\Phi(d)$$

where $d = \frac{1-c}{N-1}$ and $c \in [0, \frac{1}{N}]$. Note that $c = 0$ corresponds to $R = \log(N-1)$ and $c = \frac{1}{N}$ corresponds to $R = \log N$. By implicit differentiation, $D_h(R)$ is continuously differentiable in $(\log(N-1), \log N)$ for all $N \geq 2$, and its derivative is

$$D_h'(R_c) = \frac{\partial D_h(R_c)}{\partial c} \left(\frac{\partial R_c}{\partial c} \right)^{-1} = \frac{\rho(c/2) - \rho(d/2)}{\log(d/c)}. \tag{3.14}$$

Assume that $N \geq 2$. By L'Hospital's rule, the left derivative at $R = \log N$ is given by the limit

$$\begin{aligned} D_{h^-}'(\log N) = \lim_{R \to (\log N)^-} D_h'(R) &= \lim_{c \to 1/N} \frac{\rho(c/2) - \rho(d/2)}{\log(d/c)} \\ &= \lim_{c \to 1/N} \frac{\dfrac{\rho(c/2) - \rho(d/2)}{d/2 - c/2}}{\dfrac{\log(d/2) - \log(c/2)}{d/2 - c/2}}. \end{aligned} \tag{3.15}$$

The denominator in (3.15) converges to $2N \log e$. On the other hand, we have

$$\begin{aligned} \frac{\rho(c/2) - \rho(d/2)}{d/2 - c/2} &= \frac{\rho\left(\frac{c}{2}\right) - \rho\left(\frac{1}{2N}\right)}{\frac{c}{2} - \frac{1}{2N}} \frac{\frac{c}{2} - \frac{1}{2N}}{\frac{d}{2} - \frac{c}{2}} \\ &\quad - \frac{\rho\left(\frac{d}{2}\right) - \rho\left(\frac{1}{2N}\right)}{\frac{d}{2} - \frac{1}{2N}} \frac{\frac{d}{2} - \frac{1}{2N}}{\frac{d}{2} - \frac{c}{2}}. \end{aligned} \tag{3.16}$$

The convexity of $u(t) = \rho(e^t)$ implies that its left and right derivatives $u'_-(t)$ and $u'_+(t)$ exist for all t, which readily implies the existence of the left and right derivatives $\rho'_-(x)$ and $\rho'_+(x)$ at every $x > 0$. In fact, $u'_-(t) = e^t\rho'_-(e^t)$ and $u'_+(t) = e^t\rho'_+(e^t)$, and therefore $\rho'_-(x)$ and $\rho'_+(x)$ are positive for all $x > 0$. Thus, as $c \to \frac{1}{N}$, the first term on the right side of (3.16) converges to $-\frac{N-1}{N}\rho'_-\left(\frac{1}{2N}\right)$, and the second term converges to $-\frac{1}{N}\rho'_+\left(\frac{1}{2N}\right)$. Therefore by (3.15) the left derivative of $D_h(R)$ at $R = \log N$ is

$$D'_{h^-}(\log N) = -\frac{1}{2N^2 \log e}\rho'_+\left(\frac{1}{2N}\right) - \frac{N-1}{2N^2 \log e}\rho'_-\left(\frac{1}{2N}\right) < 0.$$

Now let $N \geq 1$. To determine the right derivative of $D_h(R)$ at $R = \log N$, replace N by $N+1$ in the range of the parameter c (so now we have $c \in (0, \frac{1}{N+1})$ and $d = \frac{1-c}{N}$). We obtain

$$D'_{h^+}(\log N) = \lim_{R \to (\log N)^+} D'_h(R) = \lim_{c \to 0} D'_h(R_c) = 0$$

since $d \to \frac{1}{N}$ as $c \to 0$.

(ii) We have $d(x, y) = |x - y|^r$ with $r > 0$, so that Theorem 3.1 and Corollary 3.1 apply. Then by (3.14), the derivative of $D_h(R)$ in $(\log(N-1), \log N)$ is parametrically given by

$$D'_h(R_c) = \frac{(c/2)^r - (d/2)^r}{\log(d/c)}$$

where $c \in (0, \frac{1}{N})$. It follows that $D_h(R)$ is twice differentiable in $(\log(N-1), \log N)$ and its second derivative is

$$\begin{aligned} D''_h(R_c) &= \frac{\partial D'_h(R_c)}{\partial c}\left(\frac{\partial R_c}{\partial c}\right)^{-1} \\ &= \frac{rc\big[(1-c)(c/2)^{r-1} + d(d/2)^{r-1}\big]\log(d/c) + 2\log e\big[(c/2)^r - (d/2)^r\big]}{2c(1-c)[\log(d/c)]^3}. \end{aligned}$$

Next we show that $D''_h(R) < 0$ in $(\log(N-1), \log N)$ for all $N \geq 2$ if $0 < r \leq 3$. By continuity, this implies that $D_h(R)$ is strictly concave on $[\log(N-1), \log N]$ for all $N \geq 2$.

Since $c < d$, $D_h''(R_c) < 0$ if and only if the numerator of the above quotient is negative. Letting $x = d/c$, after some algebra we obtain that $D_h''(R_c) < 0$ holds for all $c \in (0, \frac{1}{N})$ if and only if

$$(x^r - 1)((N-1)x + 1) - r(x^r + (N-1)x)\ln x > 0 \tag{3.17}$$

for all $x > 1$. Now observe that the left side of the preceding inequality is a linear function of $N - 1$ such that the coefficient of $N - 1$ is $x(x^r - 1 - \ln x^r)$. Since this coefficient is positive for all $x > 1$ by the inequality $t - 1 > \ln t$, $(t \neq 1)$, it is enough to prove (3.17) for $N = 2$. Equivalently, we will show that

$$f_r(x) = \frac{(x^r - 1)(x+1)}{x + x^r} - r\ln x > 0$$

for all $x > 1$. Setting $r = 3$ we have

$$f_3'(x) = \frac{(x-1)^4(x^2 + x + 1)}{(x + x^3)^2}$$

and therefore $f_3'(x) > 0$ for all $x > 1$. Since $f_3(1) = 0$, we obtain $f_3(x) > 0$ for all $x > 1$.

On the other hand,

$$\frac{\partial f_r(x)}{\partial r} = \frac{(x^r - 1)(x^2 - x^r)\ln x}{(x + x^r)^2}.$$

Hence for all $x > 1$, $\partial f_r(x)/\partial r > 0$ if $r < 2$, and $\partial f_r(x)/\partial r < 0$ if $r > 2$. Since $f_0(x) = 0$ and $f_3(x) > 0$ for all $x > 1$, this implies that $f_r(x) > 0$ for all $x > 1$ and $0 < r \leq 3$, proving the claim (ii).

Lemma 3.3 at the end of the next section shows that if $r > 3$, then there exists an $x_r > 1$ such that $f_r(x) < 0$ for all $x \in (1, x_r)$. Thus $D_h(R)$ is no longer concave on $[0, 1]$ for $r > 3$. □

3.2 Lemmas

Lemma 3.1 *If $R > 0$ and $\log n > R$, the function $\delta^*(l, n, R)$ defined in (3.11) is a strictly increasing function of l for $l_{\min} \le l \le n-1$.*

Proof. Although $\delta^*(l, n, R)$ has been defined for integer l, the defining formulas clearly allow any real $l \in (l_{\min} - \epsilon, n - 1 + \epsilon)$ for $\epsilon > 0$ small enough. We will show that in this interval, $\frac{\partial \delta^*}{\partial l} > 0$.

Since $h(c, l, n)$ has continuous partial derivatives with respect to c and l, and $\frac{\partial h}{\partial c} > 0$ (see (3.10)), the implicit function theorem implies that the partial derivative of $c(l, n, R)$ with respect to l is

$$\frac{\partial c}{\partial l} = -\frac{\partial h}{\partial l}\left(\frac{\partial h}{\partial c}\right)^{-1}.$$

Now since $\delta^*(l, n, R) = \delta(c(l, n, R), l, n)$, the chain rule gives

$$\frac{\partial \delta^*}{\partial l} = \frac{\partial \delta}{\partial c}\frac{\partial c}{\partial l} + \frac{\partial \delta}{\partial l}.$$

The partial derivatives are

$$\frac{\partial h}{\partial l} = c(\log d - \log c) - (d-c)\log e, \qquad \frac{\partial h}{\partial c} = l(\log d - \log c)$$

and

$$\frac{\partial \delta}{\partial l} = \Phi(c) - \Phi(d) + (d-c)\rho\left(\frac{d}{2}\right), \qquad \frac{\partial \delta}{\partial c} = l\left(\rho\left(\frac{c}{2}\right) - \rho\left(\frac{d}{2}\right)\right).$$

Therefore

$$\frac{\partial \delta^*}{\partial l} = \Phi(c) - \Phi(d) + (d-c)\rho\left(\frac{d}{2}\right) - \frac{\big((d-c)\log e - c\log(d/c)\big)\big(\rho\left(d/2\right) - \rho\left(c/2\right)\big)}{\log\left(d/c\right)}.$$

Since $d = \frac{1-cl}{n-l} > c$, we have $\frac{\partial \delta^*}{\partial l} > 0$ if and only if

$$\left[\Phi(c) - \Phi(d) + d\rho\left(\frac{d}{2}\right) - c\rho\left(\frac{c}{2}\right)\right]\ln\left(\frac{d}{c}\right) > (d-c)\left[\rho\left(\frac{d}{2}\right) - \rho\left(\frac{c}{2}\right)\right]. \quad (3.18)$$

In the rest of the proof we will show that (3.18) holds for all $d > c > 0$ which will imply the claim of the lemma.

By assumption, $u(t) = \rho(e^t)$ is convex, and hence absolutely continuous, i.e., it is the integral of its derivative which exists almost everywhere (see, e.g., [31]). It follows that $\rho(x) = u(\ln x)$ is also absolutely continuous, and since its right derivative $\rho'_+(x)$ exists for all $x > 0$ ($\rho'_+(x) = x^{-1}u'_+(\ln x)$), we have

$$\rho\left(\frac{d}{2}\right) - \rho\left(\frac{c}{2}\right) = \int_{c/2}^{d/2} \rho'_+(x)\,dx. \tag{3.19}$$

If $\rho(x)$ is differentiable at some $x > 0$, then $x\rho(x/2) - \Phi(x)$ is also differentiable at this point and

$$\frac{d}{dx}\left[x\rho\left(\frac{x}{2}\right) - \Phi(x)\right] = \frac{x}{2}\rho'_+\left(\frac{x}{2}\right).$$

The absolute continuity of ρ implies that $x\rho(x/2)$ is also absolutely continuous, and since Φ is continuously differentiable, we obtain that

$$\begin{aligned}\Phi(c) - \Phi(d) + d\rho\left(\frac{d}{2}\right) - c\rho\left(\frac{c}{2}\right) &= \int_c^d \frac{x}{2}\rho'_+\left(\frac{x}{2}\right)\,dx \\ &= 2\int_{c/2}^{d/2} x\rho'_+(x)\,dx. \end{aligned}\tag{3.20}$$

This and (3.19) show that (3.18) can be rewritten as

$$\left(\frac{2}{d-c}\int_{c/2}^{d/2} x\rho'_+(x)\,dx\right)\left(\frac{2}{d-c}\int_{c/2}^{d/2} \frac{1}{x}\,dx\right) > \left(\frac{2}{d-c}\int_{c/2}^{d/2} \rho'_+(x)\,dx\right).$$

Letting $f(x) = x\rho'_+(x)$ and $g(x) = 1/x$, the above is equivalent to

$$E\{f(Z)\}E\{g(Z)\} > E\{f(Z)g(Z)\} \tag{3.21}$$

where Z is a random variable uniformly distributed over the interval $(c/2, d/2)$. Since $f(x) = u'_+(\ln x)$, the strict convexity of u implies that f is strictly increasing. Since

g is strictly decreasing, (3.21) holds by a classical inequality of Chebyshev [16]. (In fact, (3.21) easily follows by expanding the expectation

$$E\big[\big(f(Z)-f(Y)\big)\big(g(Z)-g(Y)\big)\big]$$

where Y is independent of Z but has the same distribution, and by noticing that the expectation is negative since $\big(f(x)-f(y)\big)\big(g(x)-g(y)\big)<0$ for all $x,y>0$, $x\neq y$.)

$\square$

Lemma 3.2 *Let $R>0$ and let N be a positive integer such that $R<\log N$. If $n\geq N$, then $\delta^*(n-N+1,n,R)$ defined in (3.12) is a strictly increasing function of n.*

Proof. Just as in Lemma 3.1, the definition of $\delta^*(n-N+1,n,R)$ can be extended to any real valued n such that $n\geq N-\epsilon$ (if $\epsilon>0$ is small enough). We will show that $\frac{\partial\delta^*(n-N+1,n,R)}{\partial n}>0$ for all $n\geq N$.

To simplify the notation, let

$$\hat{\delta}(c,n)=(n-N+1)\Phi(c)+(N-1)\Phi(d)$$

and

$$\hat{h}(c,n)=-(n-N+1)c\log c-(1-(n-N+1)c)\log d.$$

Then the chain rule implies

$$\frac{\partial\delta^*(n-N+1,n,R)}{\partial n}=\frac{\partial\hat{\delta}}{\partial c}\frac{\partial c}{\partial n}+\frac{\partial\hat{\delta}}{\partial n}. \tag{3.22}$$

We have

$$\begin{aligned}\frac{\partial\hat{\delta}}{\partial c} &= (n-N+1)\left(\rho\left(\frac{c}{2}\right)-\rho\left(\frac{d}{2}\right)\right)\\ \frac{\partial\hat{\delta}}{\partial n} &= \Phi(c)-c\rho\left(\frac{d}{2}\right)\end{aligned}$$

and by implicit differentiation in (3.13),

$$\frac{\partial c}{\partial n} = -\frac{\partial \hat{h}}{\partial n}\left(\frac{\partial \hat{h}}{\partial c}\right)^{-1} = -\frac{c(\log d - \log c + \log e)}{(n-N+1)(\log d - \log c)}.$$

Substitution into (3.22) gives

$$\frac{\partial \delta^*(n-N+1,n,R)}{\partial n} = \Phi(c) - c\rho\left(\frac{c}{2}\right) + \frac{c\big(\rho(d/2)-\rho(c/2)\big)}{\ln(d/c)}.$$

Since $d = \frac{1-c(n-N+1)}{N-1} > c$ if $c \in (0, 1/n)$, we have $\frac{\partial \delta^*(n-N+1,n,R)}{\partial n} > 0$ if and only if

$$\frac{\rho\left(d/2\right)-\rho\left(c/2\right)}{\ln(d/2)-\ln(c/2)} > \rho\left(\frac{c}{2}\right) - \frac{\Phi(c)}{c}. \tag{3.23}$$

By (3.20), the right side of the preceding inequality equals

$$\frac{1}{c/2}\int_0^{c/2} x\rho'_+(x)\,dx = \frac{1}{c/2}\int_0^{c/2} u'_+(\ln x)\,dx$$

where $u(t) = \rho(e^t)$. Since u'_+ is strictly increasing, the last expression is less than $u'_+(\ln(c/2))$. On the other hand, since u is strictly convex,

$$\frac{\rho\left(d/2\right)-\rho\left(c/2\right)}{\ln(d/2)-\ln(c/2)} = \frac{u(\ln(d/2))-u(\ln(c/2))}{\ln(d/2)-\ln(c/2)} > u'_+(\ln(c/2))$$

which proves (3.23). □

Lemma 3.3 *Let*

$$f_r(x) = \frac{(x^r-1)(x+1)}{x+x^r} - r\ln x.$$

If $r > 3$, then there is an $x_r > 1$ such that $f_r(x) < 0$ for all $x \in (1, x_r)$.

Proof. Let $r > 3$. Since $f_r(1) = 0$, it is enough to prove that $f'_r(x) < 0$ in a nonempty interval $(1, x_r)$. We have

$$f'_r(x) = \frac{x^{2r} - rx^{2r-1} + rx^{r+1} - 2x^r + rx^{r-1} - rx + 1}{(x+x^r)^2} = \frac{g_r(x)}{(x+x^r)^2}.$$

Since $g_r(1) = 0$, it is sufficient to prove that $g_r'(x) < 0$ in some $(1, x_r)$. We have

$$g_r'(x) = 2rx^{2r-1} - r(2r-1)x^{2r-2} + r(r+1)x^r - 2rx^{r-1} + r(r-1)x^{r-2} - r.$$

Since $g_r'(1) = 0$, it is enough to prove that $g_r''(x) < 0$ in some $(1, x_r)$. We have

$$\begin{aligned} g_r''(x) &= rx^{r-3}\big(2(2r-1)x^{r+1} - (2r-1)(2r-2)x^r \\ &\quad + r(r+1)x^2 - 2(r-1)x + (r-1)(r-2)\big) \\ &= rx^{r-3}h_r(x). \end{aligned}$$

Since $h_r(1) = 2r(3-r) < 0$ if $r > 3$, there exists an $x_r > 1$ such that $f_r(x) < 0$ for all $x \in (1, x_r)$, as claimed. $\square$

Chapter 4

The structure and existence of optimal ECSQs for nonatomic distributions

As we have discussed in Section 2.3, the nearest neighbor condition implies that optimal fixed-rate quantizers can be assumed to be regular. This fact reveals an important structural property of optimal fixed-rate quantizers and it leads to efficient design and simple implementation procedures (see Section 2.2). Moreover, regularity allows the use of efficient tools in proving the existence and exploring the properties of optimal quantizers.

On the other hand, not much is known about the optimal entropy-constrained quantization even in the scalar case, although it appears that the regularity and existence of optimal entropy constrained quantizers is implicitly assumed in the literature. In this chapter the problem of regularity and existence of optimal ECSQs are addressed for continuous sources and convex difference distortion measures.

In Section 4.1 we prove regularity type properties of entropy-constrained scalar quantizers. Theorem 4.1 shows that for any finite-point quantizer there is a regular quantizer with the same entropy and not greater distortion. As a consequence of this fact, Corollary 4.2 shows that regular finite-point quantizers can perform arbitrarily close to the operational distortion-rate curve $D_h(R)$. Theorem 4.2 extends Theorem 4.1 to infinite-point quantizers; it shows that any quantizer can be replaced with an almost regular quantizer, that is, with a quantizer which has interval cells but may be undefined on a set of probability zero.

In Section 4.2 the existence of optimal ECSQs are proved. Using the results of Section 4.1, we prove in Corollary 4.3 that for any fixed N, a regular optimal ECSQ exists among all quantizers having at most N codepoints. Theorem 4.3 shows that optimal ECSQs always exist, and that such optimal quantizers can be assumed to be almost regular.

In Section 4.3 we show how to strengthen these results by specializing to mean squared error. Here we consider sources with densities which satisfy certain regularity properties; however the resulting class of densities still contains all source densities commonly used in parametric source models. The main result of the section is Theorem 4.5, where we prove the existence of an optimal ECSQ which is regular. In Lemma 4.3 (which is in fact a tool in the proof) the modified nearest neighbor condition (2.3) is extended to regular optimal quantizers not on the lower convex hull of $D_h(R)$.

Finally, a discussion of regularity-type properties is given in Section 4.4.

Throughout this chapter we will assume that d is a difference distortion measure of the form

$$d(x, y) = \rho(|x - y|) \tag{4.1}$$

where $\rho : [0, \infty) \to [0, \infty)$ is a nondecreasing function. Furthermore, without loss of generality we assume that $\rho(0) = 0$. Also, we assume that

$$E\{\rho(|X - c|)\} < \infty \qquad \text{for some } c \in \mathbb{R}. \tag{4.2}$$

Lemma A.3 in the Appendix shows that if ρ is also *lower semicontinuous* (that is, $\liminf_{x\to x_0} \rho(x) \geq \rho(x_0)$ for all $x_0 \geq 0$), then for any Borel set $S \subset \mathbb{R}$ there is a $c^* \in \mathbb{R}$ (the "generalized centroid" of S) minimizing the distortion over S, i.e.,

$$\int_S d(x, c^*) d\mu(x) = \inf_{c\in\mathbb{R}} \int_S d(x, c) d\mu(x) < \infty. \tag{4.3}$$

Since ρ is nondecreasing, if S is an interval, then (at least one of) the minimizing c^* lies inside S. It can also be shown that if ρ is strictly convex and satisfies (4.2), then c^* is unique for any S such that $\mu(S) > 0$ [12].

To avoid technical difficulties, we also assume that ρ satisfies the following stronger condition:

$$E\{\rho(|X - c|)\} < \infty \qquad \text{for all } c \in \mathbb{R}. \tag{4.4}$$

We note that all the results still hold if we only assume that the weaker condition (4.2) holds; however some of the proofs become more technical.

For example, (4.4) holds if

$$E\{\rho(|X|)\} < \infty, \quad \text{and} \quad \limsup_{x\to\infty} \frac{\rho(x+a)}{\rho(x)} < \infty \text{ for all } a > 0.$$

In case of the *mean squared error* (MSE), both (4.2) and (4.4) are equivalent to the familiar condition $E\{X^2\} < \infty$.

4.1 Optimality vs. regularity

For difference distortion measures in the form of (4.1), nearest neighbor quantizers (i.e., quantizers which satisfy $d(x, Q(x)) = \min_{1\leq i\leq N} d(x, c_i)$ for all x) are regular, and thus – as we have already seen – an optimal fixed-rate N-point quantizer can be assumed to be regular.

The situation is different in the entropy-constrained case. An optimal ECSQ is not necessarily a nearest neighbor quantizer, and thus in general it is incorrect to restrict the attention to regular quantizers (or quantizers with interval cells) when searching for an optimal ECSQ. Indeed, the following example shows that for discrete sources and certain rate constraints there may not exist a regular ECSQ which achieves $D_h(R)$.

Example 4.1 Let $d(x,y) = (x-y)^2$, and let X be a discrete random variable taking values in $\{-1, 0, 2\}$ with the following distribution:

$$P\{X=-1\} = P\{X=2\} = 2/5 \quad \text{and} \quad P\{X=0\} = 1/5.$$

A quantizer Q for X is in effect defined by a partition of the set $\{-1, 0, 2\}$ by the cells of Q (the definition of Q for other values is immaterial) and by the corresponding codepoints (the optimal codepoints for each partition can easily be found). Checking the five possible partitions, it turns out that for all R such that

$$h_b(1/5) \leq R < h_b(2/5)$$

where $h_b(x) = -x \log x - (1-x)\log(1-x)$, an optimal ECSQ has two codepoints $c_1 = 1/2$ and $c_2 = 0$, and the corresponding cells must satisfy

$$\{-1, 2\} \subset S_1 \quad \text{and} \quad \{0\} \subset S_2.$$

It follows that S_1 and S_2 cannot be disjoint intervals, and thus no regular quantizer can be optimal in this case. □

The first result of this section shows that such a pathological example cannot exist if X has a nonatomic distribution (that is, when $P\{X = x\} = 0$ for all $x \in \mathbb{R}$). More specifically, we will show that if X has a nonatomic distribution, then, for non-decreasing convex difference distortion measures, just as in fixed-rate quantization, every finite-point ECSQ can be "regularized" so that the resulting regular quantizer has the same cell probabilities (and hence the same entropy) and equal or less distortion. The key to this result is the following lemma.

Lemma 4.1 *Assume that the random variable X has a nonatomic distribution μ supported in the interval (σ, τ), and let $d(x, y) = \rho(|x - y|)$ where $\rho : [0, \infty) \to [0, \infty)$ is a convex and nondecreasing function. Let Q be an arbitrary two-point quantizer with cells S_1, S_2 and corresponding codepoints c_1, c_2 such that $c_1 < c_2$. Then there exists a two-point quantizer $\widehat{Q}$ with interval cells $\widehat{S}_1, \widehat{S}_2$ and corresponding codepoints c_1, c_2 such that $\widehat{S}_1 < \widehat{S}_2$,*[1] *$\mu(S_i) = \mu(\widehat{S}_i)$, $i = 1, 2$, and $D(\widehat{Q}) \le D(Q)$.*

Remark. $\widehat{Q}$ has interval cells but it is not necessarily regular ($c_i \in \widehat{S}_i$ is not guaranteed by the construction). However, since ρ is nondecreasing, the centroid rule (4.3) guarantees that each c_i can be replaced by a $\hat{c}_i \in \widehat{S}_i$ such that the distortion is not increased.

Proof. By the continuity of the distribution function of μ, there is an $r \in (\sigma, \tau)$ such that $\mu((\sigma, r]) = \mu(S_1)$. Let $\widehat{Q}$ be the quantizer with decision regions $S_1 = (\sigma, r]$, $S_2 = (r, \tau)$, and codepoints $\{c_1, c_2\}$. Then $H(\widehat{Q}) = H(Q)$. Next we show that $D(\widehat{Q}) \le D(Q)$. The key observation is that the function

$$\phi(x) = \rho(|x - c_1|) - \rho(|x - c_2|)$$

[1] Recall that for $A, B \subset \mathbb{R}$, the notation $A < B$ means that $a < b$ for all $a \in A$ and $b \in B$.

is nondecreasing. To show this, rewrite ϕ in the following form

$$\phi(x) = \begin{cases} \rho(c_1 - x) - \rho(c_2 - x) & \text{if } x < c_1 \\ \rho(x - c_1) - \rho(c_2 - x) & \text{if } c_1 \le x \le c_2 \\ \rho(x - c_1) - \rho(x - c_2) & \text{if } x > c_2 \end{cases}$$

Since ρ is nondecreasing, ϕ is clearly nondecreasing in the interval $[c_1, c_2]$, and the convexity of ρ readily implies that ϕ is also nondecreasing in $(\sigma, c_1]$ and $[c_2, \tau)$.

Since $\mu(\widehat{S}_1 \cap S_2) = \mu(S_1 \cap \widehat{S}_2)$ and $\widehat{S}_1 < \widehat{S}_2$,

$$\int_{\widehat{S}_1 \cap S_2} \phi(x) d\mu(x) \le \int_{S_1 \cap \widehat{S}_2} \phi(x) d\mu(x). \tag{4.5}$$

Adding

$$\int_{S_1 \cap \widehat{S}_1} \rho(|x - c_1|) d\mu(x) + \int_{S_2 \cap \widehat{S}_2} \rho(|x - c_2|) d\mu(x)$$

to both sides and rearranging terms yields $D(\widehat{Q}) \le D(Q)$. □

We showed that for convex and nondecreasing distortion measures and for any nonatomic distribution a two-point quantizer can be replaced by a regular two-point quantizer with the same entropy and less than or equal distortion. In what follows we will call these distortion measures two-point regular. More generally, we have the following definition.

Definition 4.1 *We say that a distortion measure d is* N-point regular *for some positive integer N, if for any nonatomic distribution μ and any N-point quantizer Q there is a* regular *N-point quantizer $\widehat{Q}$ such that $D(\widehat{Q}, \mu) \le D(Q, \mu)$ with Q and $\widehat{Q}$ having the same cell probabilities according to μ. In other words, $\widehat{Q}$ should be such that there is a one-to-one mapping π between the cells of Q and $\widehat{Q}$ satisfying $\mu(S) = \mu(\pi(S))$ for all cells S of Q. Moreover, d is called* finitely regular *if it is N-point regular for every positive integer N.*

Lemma 4.1 showed that nondecreasing and convex difference distortion measures are two-point regular. The next theorem shows that if d is two-point regular, then it is also finitely regular.

Theorem 4.1 *Every two-point regular distortion measure is finitely regular. That is, if X has a nonatomic distribution μ, and d is two-point regular, then for any finite-point quantizer Q with cells $S_1, \dots, S_N$ there exists a regular quantizer $\widehat{Q}$ with interval cells $\widehat{S}_1, \dots, \widehat{S}_N$ such that $\mu(S_i) = \mu(\widehat{S}_i)$, $i = 1, \dots, N$, and $D(\widehat{Q}) \leq D(Q)$.*

Proof. The proof is done by induction on N. A one-point quantizer is always regular; hence d is always one-point regular. Also, d is two-point regular by assumption. Now assume that d is n-point regular for all $n \leq N-1$, where $N \geq 3$. We show that then d is also N-point regular.

Let $\{c_1, \dots, c_N\}$ denote the codepoints of Q, and assume without loss of generality that the indexing of the cells is such that

$$\mu(S_1) = \min_{1 \leq i \leq N} \mu(S_i). \tag{4.6}$$

The main idea in constructing $\widehat{Q}$ is the following: First we fix S_1 and, using the induction hypothesis, redefine the cells $S_2, \dots, S_N$ such that the new cells will be "intervals" in $(\sigma, \tau) \setminus S_1$, the complement of S_1. That is, the new cells will satisfy $S_1^{(1)} = S_1$ and $S_2^{(1)} < \cdots < S_N^{(1)}$. Then we redefine $S_1^{(1)}$ and $S_2^{(1)}$ such that the quantizer thus obtained have a leftmost cell which is a proper interval. Finally, the other $N-1$ cells are replaced by intervals using the induction hypothesis.

Step 1. Let μ_1 be the restriction of μ to $(\sigma, \tau) \setminus S_1$, that is, $\mu_1(B) = \mu(B \cap ((\sigma, \tau) \setminus S_1))$ for any Borel set B. Furthermore, let the $(N-1)$-point quantizer q_1 be given by the cells $\{S_1 \cup S_2, S_3, \dots, S_N\}$ and codepoints $\{c_2, \dots, c_N\}$. Then

$$D(Q, \mu) = D_1(Q, \mu) + D(q_1, \mu_1). \tag{4.7}$$

Since μ_1 is nonatomic, the induction hypothesis implies that there is a regular quantizer $\hat{q}_1$ with interval cells $T_1^{(1)} < \cdots < T_{N-1}^{(1)}$ and codepoints $\{e_1^{(1)}, \ldots, e_{N-1}^{(1)}\}$ such that

$$D(\hat{q}_1, \mu_1) \le D(q_1, \mu_1) \tag{4.8}$$

and the cells of q_1 and $\hat{q}_1$ have the same measures according to μ_1. Now define the quantizer Q_1 by the cells $\{S_1^{(1)}, \ldots, S_N^{(1)}\}$ and codepoints $\{c^{(1)}, \ldots, c_N^{(1)}\}$ such that

$$S_i^{(1)} = \begin{cases} S_1 & \text{if } i = 1 \\ T_{i-1}^{(1)} \cap ((\sigma, \tau) \setminus S_1) & \text{if } 2 \le i \le N \end{cases} \quad \text{and} \quad c_i^{(1)} = \begin{cases} c_1 & \text{if } i = 1 \\ e_{i-1}^{(1)} & \text{if } 2 \le i \le N. \end{cases}$$

From the construction of Q_1 it follows that Q and Q_1 have the same cell probabilities according to μ,

$$S_2^{(1)} < \cdots < S_N^{(1)} \tag{4.9}$$

and

$$D_i(Q_1, \mu) = \begin{cases} D_1(Q, \mu) & \text{if } i = 1 \\ D_{i-1}(\hat{q}_1, \mu_1) & \text{if } 2 \le i \le N. \end{cases}$$

Thus (4.7) and (4.8) imply that

$$D(Q_1, \mu) \le D(Q, \mu).$$

Step 2. Let μ_2 be the restriction of μ to $S_1^{(1)} \cup S_2^{(1)}$, and let the two-point quantizer q_2 have cells $\{S_1^{(1)}, (\sigma, \tau) \setminus S_1^{(1)}\}$ and codepoints $\{c_1^{(1)}, c_2^{(1)}\}$. Then, since d is assumed to be two-point regular, there is a regular quantizer $\hat{q}_2$ with interval cells $T_1^{(2)} < T_2^{(2)}$ and codepoints $\{e_1^{(2)}, e_2^{(2)}\}$ such that $D(\hat{q}_2, \mu_2) \le D(q_2, \mu_2)$ and $\hat{q}_2$ and q_2 have the same cell measures according to μ_2. We assume (as we may without loss of generality) that $T_1^{(2)} = (\sigma, t)$ and $T_2^{(2)} = [t, \tau)$, where $t = \inf\{x : \mu_2((\sigma, x)) = \mu_2(T_1^{(2)})\}$.

Since either $\mu_2(T_1^{(2)}) = \mu(S_1^{(1)})$ or $\mu_2(T_1^{(2)}) = \mu(S_2^{(1)})$, by (4.6) we have $\mu_2(T_1^{(2)}) \leq \mu(S_2^{(1)})$. Thus $t \leq \sup S_2^{(1)}$, because $S_2^{(1)} \subset (\sigma, \sup S_2^{(1)}]$, and so

$$\mu_2((\sigma, \sup S_2^{(1)}]) \geq \mu_2(S_2^{(1)}) = \mu(S_2^{(1)}) \geq \mu_2(T_1^{(2)}) = \mu_2((\sigma, t]).$$

Therefore we can define the quantizer Q_2 with cells and codepoints

$$S_i^{(2)} = \begin{cases} (\sigma, t) & \text{if } i = 1 \\ [t, \tau) \cap (S_1^{(1)} \cup S_2^{(1)}) & \text{if } i = 2 \\ S_i^{(1)} & \text{if } 3 \leq i \leq N \end{cases} \quad \text{and} \quad c_i^{(2)} = \begin{cases} e_1^{(2)} & \text{if } i = 1 \\ e_2^{(2)} & \text{if } i = 2 \\ c_i^{(1)} & \text{if } 3 \leq i \leq N. \end{cases}$$

Then, as in *Step 1*, it is easy to see that

$$D(Q_2, \mu) \leq D(Q_1, \mu)$$

and Q_1 and Q_2 have the same cell probabilities according to μ.

Step 3. Since $(\sigma, \tau) \setminus S_1^{(2)} = [t, \tau)$ is an interval, using the restriction of μ to $[t, \tau)$, by the induction hypothesis we can replace the cells $S_2^{(2)}, \cdots, S_N^{(2)}$ with interval cells $\widehat{S}_2 < \cdots < \widehat{S}_N$ and corresponding codepoints to obtain a regular quantizer $\widehat{Q}$ (where $\widehat{S}_1 = S_1^{(2)}$) such that $\widehat{Q}$ has the same cell probabilities as Q_2 according to μ, and

$$D(\widehat{Q}, \mu) \leq D(Q_2, \mu).$$

Consequently, $D(\widehat{Q}, \mu) \leq D(Q, \mu)$ and $\widehat{Q}$ and Q have the same cell probabilities. $\square$

By Lemma 4.1, the conditions of Theorem 4.1 are satisfied for $d(x, y) = \rho(|x - y|)$ with convex and nondecreasing ρ. Moreover, the construction of Lemma 4.1 does not change the codepoints of the quantizer, and it orders the cells according to the corresponding codepoints. It is easy to see that this order preserving construction is maintained in Theorem 4.1 if Lemma 4.1 is used in each step of the induction argument. Thus we obtain the following corollary.

Corollary 4.1 *Assume that the real random variable X has a nonatomic distribution μ, and let $d(x,y) = \rho(|x-y|)$ where $\rho : [0,\infty) \to [0,\infty)$ is a convex and nondecreasing function. Then for any finite-point quantizer Q with cells $S_1, \ldots, S_N$ and corresponding codepoints $c_1, \ldots, c_N$ there exists a regular quantizer $\widehat{Q}$ with interval cells $\widehat{S}_1, \ldots, \widehat{S}_N$ such that $\mu(S_i) = \mu(\widehat{S}_i)$, $i = 1, \ldots, N$, $D(\widehat{Q}) \le D(Q)$, and if $c_i < c_j$, then $\widehat{S}_i < \widehat{S}_j$.*

Another easy consequence of Theorem 4.1 is that $D_h(R)$ can be arbitrarily approached by the distortion of regular finite-point ECSQs with entropy not exceeding R. This fact can be very useful in analyzing the high rate asymptotic behavior of $D_h(R)$ [23].

Corollary 4.2 *Assume X has a nonatomic distribution μ supported in the interval (σ, τ), and let d be a two-point regular distortion measure of the form $d(x,y) = \rho(|x-y|)$ where $\rho : [0,\infty) \to [0,\infty)$ is a nondecreasing function. Then for any $R \ge 0$ and $\epsilon > 0$ there is a regular finite-point quantizer Q_ϵ such that $H(Q_\epsilon) \le R$ and*

$$D(Q_\epsilon) < D_h(R) + \epsilon.$$

Remark. By Lemma 4.1, the squared error distortion case is also included in the corollary.

Proof. First we show that for any quantizer Q with cells $\{S_1, S_2, \ldots\}$ and corresponding codepoints $\{c_1, c_2, \ldots\}$ there is a sequence of regular finite-point quantizers $\{Q_n^*\}$ such that $H(Q_n^*) \le H(Q)$ for all n and

$$\limsup_{n\to\infty} D(Q_n^*) \le D(Q). \tag{4.10}$$

To see this, let π be a permutation of the positive integers such that $\mu(S_{\pi(i)}) \ge \mu(S_{\pi(j)})$ whenever $i < j$. Now let Q_n have cells

$$S_{\pi(1)}, \ldots, S_{\pi(n-1)}, \bigcup_{i \ge n} S_{\pi(i)}$$

and corresponding codepoints $c_1, \ldots, c_{n-1}, 0$. It is clear that $H(Q_n) \leq H(Q)$. Moreover, since Q_n equals Q on the set $C_n = \bigcup_{i=1}^{n-1} S_{\pi(i)}$, and $\{C_n\}$ increases to (σ, τ), the monotone convergence theorem (see Theorem A.3 in the Appendix) and $E\{\rho(|X|)\} < \infty$ imply that

$$\lim_{n\to\infty} D(Q_n) = D(Q). \tag{4.11}$$

The proof of the above limit is given in Section A.4 in the Appendix. Since each Q_n is an n-point quantizer, by Theorem 4.1 there is a sequence of regular quantizers $\{Q_n^*\}$ such that Q_n^* is an n-point quantizer, $H(Q_n^*) = H(Q_n)$, and $D(Q_n^*) \leq D(Q_n)$ (note that Q_n^* has cell probabilities $\mu(S_{\pi(1)}), \ldots, \mu(S_{\pi(n-1)}), \mu(\bigcup_{i\geq n} S_{\pi(i)})$). Hence (4.10) is proved. Now for any $\epsilon > 0$ choose a quantizer Q'_ϵ such that $H(Q'_\epsilon) \leq R$ and $D(Q'_\epsilon) - D_h(R) < \epsilon/2$. Then by (4.10) there is a regular finite-point quantizer Q_ϵ such that $H(Q_\epsilon) \leq H(Q'_\epsilon)$ and $D(Q_\epsilon) - D(Q'_\epsilon) < \epsilon/2$, which in turn satisfies $D(Q_\epsilon) < D_h(R) + \epsilon$. $\square$

Unexpected problems arise if one wants to extend Theorem 4.1, and more specifically, Corollary 4.1 to infinite-point quantizers. Because of the construction used in Lemma 4.1, the construction of Theorem 4.1 for convex and nondecreasing difference distortion measures is order preserving in the sense that if c_i and c_j are two codepoints of Q such that $c_i < c_j$, then the corresponding cells of $\widehat{Q}$, $\widehat{S}_i$ and $\widehat{S}_j$, satisfy $\widehat{S}_i < \widehat{S}_j$. Now suppose we can extend Theorem 4.1 to infinite-point quantizers in a similar order preserving manner and assume that Q is an infinite-point quantizer with codepoints $\{c_i\}_{i=1}^{\infty}$ such that if $c_i < c_j$, then there exists a c_k such that $c_i < c_k < c_j$. Then the collection of interval cells $\{\widehat{S}_i\}_{i=1}^{\infty}$ of the regularized $\widehat{Q}$ will have the property that for any $\widehat{S}_i$ and $\widehat{S}_j$ such that $\widehat{S}_i < \widehat{S}_j$, there is an $\widehat{S}_k$ satisfying $\widehat{S}_i < \widehat{S}_k < \widehat{S}_j$. Clearly, the collection of intervals $\{\widehat{S}_i\}_{i=1}^{\infty}$ cannot form a partition of (σ, τ) in this case.

To deal with this problem we introduce the notion of almost regular quantizers. For a given finite measure μ we call a quantizer Q *μ-almost regular* if there exists an $S \subset (\tau, \sigma)$ with $\mu(S) = 0$ such that Q is defined on $(\sigma, \tau) \setminus S$, and every cell of Q is an interval containing the corresponding codepoint. Thus an almost regular quantizer has interval cells but it is not defined on a set of μ-measure zero. We can define $Q(x)$ arbitrarily for all $x \in S$ to make Q be defined everywhere (this clearly does not change the entropy and distortion of Q). As an example, let $S \subset [0, 1]$ be the Cantor ternary set (see Section A.5 in the Appendix), let μ be absolutely continuous with respect to the Lebesgue measure, and assume that $\mu([0, 1]) = 1$. Then $[0, 1] \setminus S$ is a union of countably many open intervals which form the cells of a μ-almost regular quantizer.

This definition can be extended similarly to higher dimensions.

Definition 4.2 *We call a vector quantizer Q μ-*almost regular* if there is a set $S \subset \mathcal{S}$ such that $\mu(S) = 0$, Q is defined on the set $\mathcal{S} \setminus S$, and Q has convex cells containing the corresponding codepoints.*[2]

To exhibit an example for a μ-almost regular quantizer in k dimensions we can use a corollary of the Vitali Covering Theorem [17]. The corollary states that disjoint closed balls contained in the k-dimensional open unit cube U can "almost cover" U in the sense that what is left out has Lebesgue measure zero. Now let μ be absolutely continuous with respect to the k-dimensional Lebesgue measure such that $\mu([0, 1]^k) = 1$. Let $B_1, B_2, \ldots$ be disjoint closed balls in $\mathbb{R}^k$ such that the set $S = [0, 1]^k \setminus \bigcup_i B_i$ has Lebesgue measure zero. Then $B_1, B_2, \ldots$ form the cells of an almost regular k-dimensional quantizer (see Figure 4.1).

The next result shows that for convex difference distortion measures, any infinite-point quantizer can be replaced by an almost regular quantizer with the same entropy

[2]Recall the notation of Chapter 2.

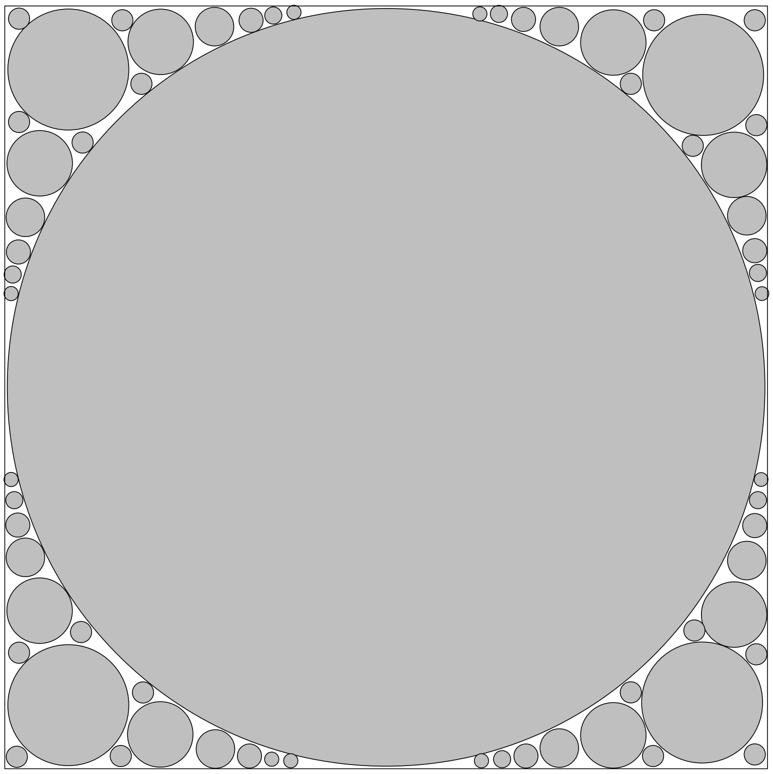

Figure 4.1: An approximation to an infinite-point almost regular vector quantizer.

and equal or less distortion.

Theorem 4.2 *Assume that the random variable X has a nonatomic distribution μ, and let $d(x,y) = \rho(|x-y|)$ where $\rho : [0,\infty) \to [0,\infty)$ is a convex and nondecreasing function. Then for any quantizer Q there exists an almost regular quantizer $\widehat{Q}$ with the same cell probabilities such that $D(\widehat{Q}) \leq D(Q)$. Moreover, if the codepoints of Q can be linearly ordered, then $\widehat{Q}$ is also regular.*

Remarks. If X has a positive density on (σ, τ), then $\widehat{Q}$ is also almost regular relative to the Lebesgue measure λ, since then μ and λ are mutually absolutely continuous with respect to each other on (σ, τ).

Note that there is an important difference in our treatments of finite- and infinite-point ECSQs. Theorem 4.2, which is valid for infinite-point quantizers, has a proof that strongly depends on the order preserving construction of Lemma 4.1, which in turn requires that ρ be a convex function. In contrast, the analogous statement for finite-point ECSQs, Corollary 4.1, is valid for all two-point regular distortion measures. We conjecture that the class of two-point regular distortion measures is larger than the class of convex and nondecreasing difference distortion measures, but so far we have not been able to find a nonconvex example for a two-point regular distortion measure.

Proof of Theorem 4.2 Special properties of convex and nondecreasing distortion measures used in Lemma 4.1 allow us to extend the proof of Theorem 4.1 to infinite-point quantizers.

If Q is a finite-point quantizer, then the statement follows from Corollary 4.1. Suppose that Q is an infinite-point quantizer with cells $\{S_1, S_2, \dots\}$ and codepoints $\{c_1, c_2, \dots\}$. The proof consists of two parts. First, for every nonatomic finite measure μ and for every positive integer n we construct an infinite-point quantizer $\widehat{Q}_n$ with $D(\widehat{Q}_n) \le D(Q)$ having cells $\{\widehat{S}_1^{(n)}, \widehat{S}_2^{(n)}, \dots\}$ and the same codepoints $\{c_1, c_2, \dots\}$ such that $\mu(S_i) = \mu(\widehat{S}_i^{(n)})$ for all i, and $c_i < c_j$ implies $\widehat{S}_i^{(n)} < \widehat{S}_j^{(n)}$ for all $i, j \le n$. Then we show that the sequence of quantizers $\widehat{Q}_n$ converges, in a sense, to a quantizer $\widehat{Q}$ defined μ-almost everywhere such that $\widehat{Q}$ has interval cells, Q and $\widehat{Q}$ have the same cell probabilities and $D(\widehat{Q}) \le D(Q)$. Finally, since ρ is nondecreasing, the centroid rule (4.3) implies that the codepoints of $\widehat{Q}$ can be redefined to obtain an almost regular quantizer such that the distortion is not increased.

The construction is rather simple with the use of Corollary 4.1. Let μ_n denote the restriction of μ to $\bigcup_{i \le n} S_i$, and apply Corollary 4.1 to μ_n and the quantizer with cells $S_1, \dots, S_n, \bigcup_{i>n} S_i$ and codepoints $c_1, \dots, c_n, 0$. The resulting quantizer

$\bar{Q}_n$ will have the same codepoints and cells $\bar{S}_1^{(n)}, \ldots, \bar{S}_{n+1}^{(n)}$ such that $\bar{S}_i^{(n)} < \bar{S}_j^{(n)}$ if $c_i < c_j$ for all $1 \le i, j \le n$, and $\mu(\bar{S}_{n+1}^{(n)} \cap (\bigcup_{i \le n} S_i)) = 0$. Define the cells of the quantizer $\widehat{Q}_n$ as $\widehat{S}_i^{(n)} = \bar{S}_i^{(n)} \cap (\bigcup_{i \le n} S_i)$ for $1 \le i \le n$, and $\widehat{S}_i^{(n)} = S_i$ for $i > n$ with corresponding codepoints $\{c_i\}$. It is easy to see that $\mu(\widehat{S}_i^{(n)}) = \mu(S_i)$ for all i, and from the construction of Corollary 4.1 it follows that

$$\begin{aligned} D(\widehat{Q}_n) &= D(\bar{Q}_n, \mu_n) + \sum_{i>n} D_i(Q, \mu) \\ &\le D(Q, \mu_n) + \sum_{i>n} D_i(Q, \mu) \\ &= D(Q). \end{aligned}$$

Thus $\widehat{Q}_n$ satisfies the required conditions for all n.

In the second part of the proof we show that there is a quantizer $\widehat{Q}$ defined μ-almost everywhere with codepoints $\{c_1, c_2, \ldots\}$ and interval cells $\{\widehat{S}_1, \widehat{S}_2, \ldots\}$ such that $\widehat{Q}_n \to \widehat{Q}$ in the sense that for all i,

$$\lim_{n\to\infty} \mu(\widehat{S}_i \triangle \widehat{S}_i^{(n)}) = 0 \tag{4.12}$$

where $\widehat{S}_i \triangle \widehat{S}_i^{(n)} = (\widehat{S}_i \setminus \widehat{S}_i^{(n)}) \cup (\widehat{S}_i^{(n)} \setminus \widehat{S}_i)$, the symmetric difference of $\widehat{S}_i$ and $\widehat{S}_i^{(n)}$. We also show that (4.12) implies

$$\lim_{n\to\infty} D_i(\widehat{Q}_n) = D_i(\widehat{Q}). \tag{4.13}$$

Then by Fatou's lemma (see Theorem A.1 in the Appendix)

$$\begin{aligned} D(Q) &\ge \liminf_{n\to\infty} D(\widehat{Q}_n) = \liminf_{n\to\infty} \sum_{i=1}^{\infty} D_i(\widehat{Q}_n) \\ &\ge \sum_{i=1}^{\infty} \liminf_{n\to\infty} D_i(\widehat{Q}_n) = \sum_{i=1}^{\infty} D_i(\widehat{Q}) \\ &= D(\widehat{Q}). \end{aligned}$$

Since $\mu(\widehat{S}_i) = \mu(\widehat{S}_i^{(n)}) = \mu(S_i)$ for all i, we have $H(\widehat{Q}) = H(Q)$, and so the proof is complete after optimizing the codepoints of $\widehat{Q}$ according to the centroid rule (4.3).

To show the existence of $\widehat{Q}$ satisfying (4.12) and (4.13), for every $i \in \mathbb{N}$ (where $\mathbb{N}$ denotes the set of positive integers) let $C_i = \{j \in \mathbb{N} : c_j < c_i\}$, and consider the sequence of sets $\{\widehat{S}_i^{(n)}\}_{n=i}^{\infty}$. From the construction of $\widehat{Q}_n$ it follows that if $n \geq i$, then $\widehat{S}_j^{(n)} \subset (\sigma, \inf \widehat{S}_i^{(n)}]$ if $j \in C_i \cap \{1, \ldots, n\}$, and $\widehat{S}_j^{(n)} \subset [\inf \widehat{S}_i^{(n)}, \tau)$ if $j \in \{1, \ldots, n\} \setminus C_i$. Since $\mu(S_i) = \mu(\widehat{S}_i^{(n)})$, we have

$$\sum_{j \in C_i \cap \{1, \ldots, n\}} \mu(S_j) \leq \mu((\sigma, \inf \widehat{S}_i^{(n)}])$$

and

$$\mu([\inf \widehat{S}_i^{(n)}, \tau)) \geq \sum_{j \in \{1, \ldots, n\} \setminus C_i} \mu(S_j). \tag{4.14}$$

On the other hand, the sequence of sets $\{C_i \cap \{1, \ldots, n\}\}_{n=i}^{\infty}$ increases to C_i, and the sequence $\{\{1, \ldots, n\} \setminus C_i\}_{n=i}^{\infty}$ increases to $\mathbb{N} \setminus C_i$, and so

$$\lim_{n\to\infty} \sum_{j \in C_i \cap \{1, \ldots, n\}} \mu(S_j) = \sum_{j \in C_i} \mu(S_j) \quad \text{and} \quad \lim_{n\to\infty} \sum_{\{1, \ldots, n\} \setminus C_i} \mu(S_j) = \sum_{j \in \mathbb{N} \setminus C_i} \mu(S_j).$$

Therefore,

$$\begin{aligned} \sum_{j \in C_i} \mu(S_j) &\leq \liminf_{n\to\infty} \mu((\sigma, \inf \widehat{S}_i^{(n)}]) \\ &= \liminf_{n\to\infty} \mu((\sigma, \inf \widehat{S}_i^{(n)})) \\ &\leq \limsup_{n\to\infty} \mu((\sigma, \inf \widehat{S}_i^{(n)})) \leq \sum_{j \in C_i} \mu(S_j) \end{aligned}$$

where the last inequality follows from (4.14). This implies

$$\lim_{n\to\infty} \mu((\sigma, \inf \widehat{S}_i^{(n)})) = \sum_{j \in C_i} \mu(S_j). \tag{4.15}$$

A similar argument shows that

$$\lim_{n\to\infty} \mu((\sigma, \sup \widehat{S}_i^{(n)})) = \sum_{j \in C_i \cup \{i\}} \mu(S_j). \tag{4.16}$$

Now let

$$\begin{aligned} a_i &= \sup\left\{x \in (\sigma,\tau) : \mu((\sigma,x)) = \sum_{j\in C_i} \mu(S_j)\right\} \\ b_i &= \inf\left\{x \in (\sigma,\tau) : \mu((\sigma,x)) = \sum_{j\in C_i\cup\{i\}} \mu(S_j)\right\} \end{aligned}$$

and define the ith cell of $\widehat{Q}$ as $\widehat{S}_i = [a_i, b_i)$ with codepoint c_i. Note that the $\widehat{S}_i$ are disjoint. Moreover, since μ is nonatomic, $\mu(\widehat{S}_i) = \mu(S_i) = \mu(\widehat{S}_i^{(n)})$, and so

$$\begin{aligned} 0 &\le \mu(\widehat{S}_i \triangle \widehat{S}_i^{(n)}) \\ &= 2\mu(\widehat{S}_i^{(n)} \setminus \widehat{S}_i) \\ &= 2\mu(\widehat{S}_i^{(n)} \setminus [a_i, b_i)) \\ &\le 2\mu([\inf \widehat{S}_i^{(n)}, a_i] \cup [b_i, \sup \widehat{S}_i^{(n)}]) \\ &\le 2\mu([\inf \widehat{S}_i^{(n)}, a_i]) + 2\mu([b_i, \sup \widehat{S}_i^{(n)}]) \end{aligned}$$

where we used the convention that if $b < a$, then $[a, b]$ stands for the empty set. By (4.15), (4.16) and the definition of a_i and b_i, the last expression goes to zero as $n \to \infty$, which proves (4.12).

Finally we prove (4.13). Clearly, for every i there is an $\alpha > 0$ such that $\mu([a_i - \alpha, b_i + \alpha] \setminus \widehat{S}_i^{(n)}) = 0$ if n is large enough. Letting $\rho_{\max} = \max_{x\in[a_i-\alpha, b_i+\alpha]} \rho(|x - c_i|)$, for such n we have

$$\begin{aligned} |D_i(\widehat{Q}_n) - D_i(\widehat{Q})| &\le \int_{\widehat{S}_i \triangle \widehat{S}_i^{(n)}} \rho(|x - c_i|)d\mu(x) \\ &\le \mu(\widehat{S}_i \triangle \widehat{S}_i^{(n)})\rho_{\max} \to 0 \end{aligned}$$

as $n \to \infty$.

If the $\{c_i\}$ can be linearly ordered, then reindexing the c_i by the set of all integers, without loss of generality we can assume that $\cdots < c_{-1} < c_0 < c_1 < \cdots$, and so

$\widehat{S}_i < \widehat{S}_{i+1}$ for all i. Then letting $a_i^* = b_{i-1}^* = (b_{i-1} + a_i)/2$ yields a regular quantizer with cells $[a_i^*, b_i^*)$ and corresponding optimal codepoints $\hat{c}_i^*$. (Since ρ is nondecreasing, by the centroid rule $\hat{c}_i^*$ can be chosen such that $\hat{c}_i^* \in [a_i^*, b_i^*)$ for all i.) This completes the proof. □

4.2 The existence of optimal ECSQs

If X has a nonatomic distribution, then by Theorem 4.1, any N-point ECSQ (where N is finite) can be replaced by a regular N-point ECSQ which has the same entropy and equal or less distortion. Since the distortion and entropy of an N-level regular ECSQ whose codepoints are optimally chosen is determined by $N-1$ real parameters (the endpoints of the interval cells), one can easily prove the following result.

Corollary 4.3 *Assume that the random variable X has a nonatomic distribution μ, and let $d(x,y) = \rho(|x-y|)$ be a two-point regular distortion measure, where $\rho : [0,\infty) \to [0,\infty)$ is a nondecreasing lower semicontinuous function. Then for any $R \geq 0$ there exists a regular quantizer with at most N points achieving*

$$D_{h,N}(R) = \inf\{D(Q) : Q \textit{ has at most } N \textit{ codepoints}, H(Q) \leq R\}.$$

Remark. The conditions of this corollary are satisfied when ρ is convex and nondecreasing, and thus Corollary 4.3 holds for the squared error distortion.

Note that the quantizer achieving $D_{h,N}(R)$ may have less than N codepoints. For the squared error distortion case and for sources with log-concave densities, Kieffer *et al.* [21] provided conditions under which $D_{h,N}$ is achieved by a quantizer having exactly N codepoints.

Proof. Let $\{Q_i\}$ be a sequence of quantizers with at most N codepoints such that $H(Q_i) \leq R$ for all i and $\lim_{i\to\infty} D(Q_i) = D_{h,N}(R)$. By Theorem 4.1, each Q_i can be assumed to be regular. Then for some $1 \leq n \leq N$, there is an infinite subsequence $\{Q_{k_i}\}$ of n-point quantizers. Because of (4.3), such a quantizer is determined by its $n-1$ ordered subdivision points, $\mathbf{q}_{k_i} = (q_1^{(k_i)}, \ldots, q_{n-1}^{(k_i)})$. Then $\{\mathbf{q}_{k_i}\}$ has a subsequence converging to a vector $\mathbf{q}$ (convergence to $-\infty$ and ∞ is allowed). Some consecutive coordinates of $\mathbf{q}$ may be equal. Let $\mathbf{q}^*$ be an at most $(n-1)$-dimensional vector obtained by dropping all but one elements of each set of equal coordinates of $\mathbf{q}$. The first and the last coordinates of $\mathbf{q}^*$ may be σ and τ, respectively. If so, then drop the corresponding coordinates to obtain $\hat{\mathbf{q}}$ which now has all its coordinates in the interval (σ, τ). Then $\hat{\mathbf{q}}$ is a vector of the subdivision points of a quantizer Q (the coordinates dropped from $\mathbf{q}$ form cells of length zero). Since X has a nonatomic distribution, the entropy is a continuous function of the subdivision point vector. On the other hand, the lower semicontinuity of ρ and Fatou's lemma imply that the distortion is a lower semicontinuous function of the subdivision points (the proof is a special case of the proof of (4.17) to be presented later which is valid also for infinite-point quantizers). Thus $\lim_{i\to\infty} H(Q_i) = H(Q)$ and $D_{h,N}(R) = \liminf_{i\to\infty} D(Q_i) \geq D(Q)$. Therefore Q satisfies $H(Q) \leq R$ and $D(Q) = D_{h,N}(R)$. $\square$

Based on Corollary 4.2, the existence of μ-almost regular optimal ECSQs can be proved for any random variable X with a nonatomic distribution μ if the distortion measure is two-point regular.

Theorem 4.3 *Assume that the random variable X has a nonatomic distribution μ, and that d is a two-point regular distortion measure of the form $d(x, y) = \rho(|x-y|)$ where $\rho : [0, \infty) \to [0, \infty)$ is a nondecreasing and lower semicontinuous function.*

Then for any $R \geq 0$ there exists a μ-almost regular quantizer Q such that $H(Q) \leq R$ and $D(Q) = D_h(R)$.

Proof. Fix $R \geq 0$. By Corollary 4.2 there is a sequence of regular finite-point quantizers $\{Q_n\}$ such that $H(Q_n) \leq R$ for all n and

$$\lim_{n\to\infty} D(Q_n) = D_h(R).$$

For each $n \in \mathbb{N}$ and $i \in \mathbb{N}$, let $S_i^{(n)} = [a_i^{(n)}, b_i^{(n)})$ denote the cell of Q_n with the ith largest probability, let $c_i^{(n)}$ denote the corresponding codepoint, and let $p_i^{(n)} = \mu(S_i^{(n)})$. In case of ties, any ordering of the equiprobable cells suffices. If Q_n has k cells, then $S_i^{(n)}$ is formally defined to be the empty set for all $i > k$. For every n, let $\mathbf{p}_n = (p_1^{(n)}, p_2^{(n)}, \dots)$. Then $\mathbf{p}_n \in C$ for all n where

$$C = \left\{ (p_1, p_2, \dots) : p_i \geq 0 \text{ for all } i \geq 1,\, p_1 \geq p_2 \geq \cdots,\, \sum_{i=1}^{\infty} p_i = 1,\, -\sum_{i=1}^{\infty} p_i \log p_i \leq R \right\}.$$

Now for any $\mathbf{p} = (p_1, p_2, \dots) \in C$ let the random variable $Y_{\mathbf{p}}$ have distribution $P\{Y_{\mathbf{p}} = n\} = p_n$ $(n \geq 1)$. Then by a result of Wyner [36, Theorem 1] we have

$$E\{\log Y_{\mathbf{p}}\} \leq -\sum_{i=1}^{\infty} p_i \log p_i \leq R.$$

Thus Markov's inequality implies

$$P\{Y_{\mathbf{p}} > n\} = P\{\log Y_{\mathbf{p}} > \log n\} \leq \frac{E\{\log Y_{\mathbf{p}}\}}{\log n} \leq \frac{R}{\log n}.$$

Since $\mathbf{p}$ was arbitrary, it follows that the family of distributions corresponding to C is tight. Therefore, by Prokhorov's theorem (see, e.g., [2]), C is compact under pointwise convergence,[3] and thus $\{\mathbf{p}_n\}$ has a pointwise convergent subsequence (denoted also

[3]This is a key observation in the proof.

by $\{\mathbf{p}_n\}$) which converges to some $\mathbf{p} = (p_1, p_2, \dots) \in C$. Now for the corresponding subsequence of quantizers also denoted by $\{Q_n\}$ form the vectors

$$\mathbf{v}_n = (a_1^{(n)}, b_1^{(n)}, c_1^{(n)}, a_2^{(n)}, b_2^{(n)}, c_2^{(n)} \dots).$$

If S_i is empty, then define $a_i = b_i = c_i = \infty$. By Cantor's diagonalization method a subsequence $\{\mathbf{v}_{n_j}\}$ converging pointwise to a vector $\mathbf{v} = (a_1, b_1, c_1, a_2, b_2, c_2 \dots)$ can be chosen (convergence to ∞ and $-\infty$ is also allowed). For this vector we can construct a quantizer with codepoints c_i and corresponding cells $S_i = [a_i, b_i)$. If $a_i = b_i$ (in this case also $c_i = a_i$ by construction), then we let S_i be the empty set. Since for every fixed n the sets $S_i^{(n)} = [a_i^{(n)}, b_i^{(n)}), i \in \mathbb{N}$ are pairwise disjoint, it is easy to see that the sets $S_i = [a_i, b_i), i \in \mathbb{N}$ are also pairwise disjoint, because if $S_i \cap S_j \neq \emptyset$ for some $i, j \in \mathbb{N}$, then for n large enough we have $S_i^{(n)} \cap S_j^{(n)} \neq \emptyset$, a contradiction. Since μ is nonatomic, $\mu(S_i) = p_i$ for all i, and hence $\sum_i \mu(S_i) = 1$. Thus the S_i are the cells of a μ-almost regular quantizer Q, and since $\mathbf{p} \in C$, we have

$$H(Q) \leq R.$$

Denote the characteristic function of a set $S \subset \mathbb{R}$ by $\mathbb{I}_S$. Since the $S_i^{(n)}$ are intervals, by the lower semicontinuity of ρ we have

$$\liminf_{n\to\infty} \mathbb{I}_{S_i^{(n)}} \rho(|x - c_i^{(n)}|) \geq \mathbb{I}_{S_i} \rho(|x - c_i|)$$

for μ-almost all x. Then by Fatou's lemma

$$
\begin{aligned}
D(Q) &= \sum_i D_i(Q) \\
&= \sum_i \int_{S_i} \rho(|x - c_i|) d\mu(x) \\
&= \sum_i \int_{\mathbb{R}} \mathbb{I}_{S_i} \rho(|x - c_i|) d\mu(x) \\
&\le \sum_i \int_{\mathbb{R}} \liminf_{n\to\infty} \mathbb{I}_{S_i^{(n)}} \rho(|x - c_i^{(n)}|) d\mu(x) \\
&\le \sum_i \liminf_{n\to\infty} \int_{\mathbb{R}} \mathbb{I}_{S_i^{(n)}} \rho(|x - c_i^{(n)}|) d\mu(x) \\
&= \sum_i \liminf_{n\to\infty} \int_{S_i^{(n)}} \rho(|x - c_i^{(n)}|) d\mu(x) \\
&\le \liminf_{n\to\infty} \sum_i \int_{S_i^{(n)}} \rho(|x - c_i^{(n)}|) d\mu(x) \\
&= \liminf_{n\to\infty} D(Q_n) = D_h(R). \qquad (4.17)
\end{aligned}
$$

Thus Q is an optimal quantizer. □

4.3 Optimality and regularity for sources with densities

In the previous section we proved under general conditions that an optimal quantizer exists for continuous sources which is almost regular. However, we can prove the stronger result that a regular optimal quantizer exists if we restrict the attention to sources with well behaved densities and the mean squared distortion.

First we treat the case where the source has a continuous and monotone density.

Theorem 4.4 *Let X have a density function which is continuous and monotone on its support, and assume that $d(x,y) = (x-y)^2$ and $E\{X^2\} < \infty$. Then for any quantizer Q there is a regular quantizer $\widehat{Q}$ such that $H(\widehat{Q}) = H(Q)$ and $D(\widehat{Q}) \le D(Q)$. In particular, for any rate constraint $R \ge 0$ there is an optimal ECSQ which is regular.*

Proof. The theorem is an immediate corollary of Theorem 4.3 and Lemma 4.2 below. $\square$

The following lemma is a generalization of a result of Kieffer *et al.* [21, Lemma A.5] where the same statement was proved for regular finite-point quantizers.

Lemma 4.2 *Let X be a random variable with a density function f such that f is continuous, positive, and nonincreasing (resp. nondecreasing) on (σ, τ), and $E\{X^2\} < \infty$. Assume that $d(x,y) = (x-y)^2$. Then for any scalar quantizer Q there exists a regular quantizer $\widehat{Q}$ with cells $\widehat{S}_1 < \widehat{S}_2 < \cdots$ (resp. $\widehat{S}_1 > \widehat{S}_2 > \cdots$) such that Q and $\widehat{Q}$ have the same cell probabilities, $P\{X \in \widehat{S}_i\}$ is nonincreasing as i increases, and $D(\widehat{Q}) \le D(Q)$.*

Proof. Assume that f is nonincreasing, and denote the cells of Q by $\{S_1, S_2, \dots\}$. By Theorem 4.2 we can assume that Q is almost regular. If Q has a finite number of cells, then it is a regular quantizer, and the lemma holds by [21, Lemma A.5]. Now assume that Q has infinitely many cells. Denote the distribution of X by μ, and let π and Q_n^* be as in the proof of Corollary 4.2. Recall that Q_n^* is a regular n-point quantizer such that it has cell probabilities $\mu(S_{\pi(1)}), \dots, \mu(S_{\pi(n-1)}), \mu(\bigcup_{i\ge n} S_{\pi(i)})$, and so $H(Q_n^*) \le H(Q)$ for all n, and moreover $\limsup_{n\to\infty} D(Q_n^*) \le D(Q)$.

Applying the lemma for Q_n^*, for each n we obtain a regular quantizer $\widehat{Q}_n$ with subdivision points $x_0^{(n)} = \sigma < x_1^{(n)} < x_2^{(n)} < \cdots$ such that $\mu([x_{i-1}^{(n)}, x_i^{(n)})) = \mu(S_{\pi(i)})$

for $i = 1, \ldots, n$ (since Q_n^* has n cells, we formally define $x_i^{(n)} = \tau$ for all $i \geq n$), and $D(\widehat{Q}_n) \leq D(Q_n^*)$. Denote the distribution function of X by F, and let F^{-1} be the inverse of F in the interval (σ, τ). (F is invertible since the density f is positive on (σ, τ).) It is easy to see that for all $i \geq 0$,

$$\lim_{n\to\infty} F(x_i^{(n)}) = \sum_{j:\pi(j)\leq i} \mu(S_{\pi(i)}).$$

Thus letting $\hat{x}_i = F^{-1}(\sum_{j:\pi(j)\leq i} \mu(S_{\pi(i)}))$, we have $x_i^{(n)} \to \hat{x}_i$ as $n \to \infty$ for all $i \geq 0$. Therefore the quantizer $\widehat{Q}$ with subdivision points $\hat{x}_0, \hat{x}_1, \ldots$ satisfies the requirements of the lemma since

$$D(\widehat{Q}) = \lim_{n\to\infty} D(\widehat{Q}_n) \leq \limsup_{n\to\infty} D(Q_n^*) \leq D(Q).$$

A similar argument can be used when f is nondecreasing. $\square$

Somewhat relaxing the strong requirement of regularity we can easily generalize Theorem 4.4 to a wide class of source densities.. We call two disjoint intervals *consecutive* if they have a common endpoint. A function $f : (\sigma, \tau) \to \mathbb{R}$ is called *piecewise monotone* (resp. *finitely piecewise monotone*) if (σ, τ) can be partitioned into countably many (resp. finitely many) intervals such that any bounded set intersects only a finite number of these intervals and f is monotone in each of these intervals. *Piecewise continuity* is similarly defined with continuity in place of monotonicity.

Later on in this section we will prove the regularity of optimal ECSQs for sources with piecewise monotone and piecewise continuous densities. The class of these densities contains all densities commonly used as parametric source models. Note that all continuous unimodal densities belong to this class. In particular, the following densities satisfy our conditions.

- Generalized Gaussian distribution with parameters $\alpha, \beta > 0$:

$$f_G(x) = \frac{\alpha\eta(\alpha,\beta)}{2\Gamma(1/\alpha)} e^{-(\eta(\alpha,\beta)|x|)^\alpha}$$

  where $\eta(\alpha,\beta) = \beta^{-1}\sqrt{\Gamma(3/\alpha)/\Gamma(1/\alpha)}$ and Γ is the gamma function, that is, $\Gamma(r) = \int_0^\infty t^{r-1}e^{-t}dt$. This family contains, for example, the zero mean Gaussian distribution with variance β^2 ($\alpha = 2$), and the Laplacian distribution with parameter λ ($\alpha = 1, \beta = \sqrt{2}/\lambda$).

- Gamma distribution with parameters $\lambda, r > 0$:

$$f_\gamma(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{r-1}}{\Gamma(r)} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

  This family contains, for example, the exponential distribution with parameter λ ($r = 1$), and the Chi-squared distribution with n degrees of freedom ($\lambda = 1/2, r = n/2$).

- Beta distribution with parameters $\alpha, \beta > 0$:

$$f_\beta(x) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

  where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

The next corollary states that for piecewise continuous and piecewise monotone densities the optimal quantizer consists of a sequence of consecutive interval cells.

Corollary 4.4 *Let X be a random variable with a density f which is piecewise continuous and piecewise monotone in (σ,τ). Assume that $d(x,y) = (x-y)^2$ and $E\{X^2\} < \infty$. Then for any rate constraint $R \geq 0$ there is an optimal ECSQ which has interval cells and is defined everywhere in (σ,τ) except on a set $S \subset (\sigma,\tau)$ such that $S \cap (a,b)$ is a finite set for any bounded interval (a,b).*

Proof. Denote the distribution of X by μ. By Theorem 4.3 for any rate constraint $R \geq 0$ there is a μ-almost regular quantizer Q which is optimal, and we can assume that each cell of Q is of the form $[a, b)$ for some $a < b$. Since f is piecewise monotone and piecewise continuous, it is easy to see that there is an interval partition $\{I_i\}$ of (σ, τ) such that only finitely many of the I_i intersect any bounded subset of (σ, τ), and f is continuous and monotone in the interior of each I_i. Let A_i and B_i denote the cells of Q covering $\inf I_i$ and $\sup I_i$, respectively. If Q is not defined at any of these two points, then the corresponding set A_i or B_i is defined to be the empty set. Let $\hat{I}_i = I_i \setminus (A_i \cup B_i)$. Note that $\hat{I}_i$ is either empty or it is an interval, only finitely many of the $\hat{I}_i$ intersect any bounded subset of (σ, τ), and f is continuous and monotone in the interior of $\hat{I}_i$ for all i. Then by Lemma 4.2 we can assume that for all i the cells of Q contained in $\hat{I}_i$ form a sequence of consecutive intervals. These cells either cover the interior of $\hat{I}_i$, or what is left out is an interval of probability zero which can be assumed to be a cell of Q. Thus the set of points where Q may not be defined is a subset of the set $S = \{\inf \hat{I}_i, \sup \hat{I}_i;\, i = \ldots, -1, 0, 1, \ldots\}$. From the construction of the sets $\{\hat{I}_i\}$ it is easy to see that $S \cap (a, b)$ is a finite set for any bounded interval (a, b). □

To prove regularity under the conditions of Corollary 4.4 is much harder, and to do it we need an additional tool provided by a result due to Chou and Betts [4]. They showed that if an optimal entropy constrained (scalar or vector) quantizer achieves the lower convex hull of $D_h(R)$, then it satisfies the modified nearest neighbor condition (see Section A.2 in the Appendix). It is not hard to see that quantizers satisfying the modified nearest neighbor condition are regular. This result is very general in that it does not require the source distribution to be nonatomic and is also valid for vector quantizers. On the other hand, its assumptions (namely, the existence of an optimal entropy-constrained quantizer which achieves the lower convex hull of $D_h(R)$) are very hard to check since little is known about the properties of $D_h(R)$. Also, for the uniform distribution, one of the very few cases where $D_h(R)$ is analytically known, it was shown in Chapter 3 that there exist only isolated rate values at which the lower convex hull of $D_h(R)$ is achieved by an optimal ECSQ.

Although the direct applicability of [4] seems to be limited, its proof technique turns out to be very useful in our investigation. Following the ideas of [4], the next lemma, which extends the validity of the modified nearest neighbor condition to all optimal regular ECSQs, plays a key role in establishing the regularity of the optimal ECSQs of Corollary 4.4.

Lemma 4.3 *Let X be a random variable with a positive and continuous density f in the interval (σ, τ). Let $d(x, y) = \rho(|x-y|)$, where $\rho : [0, \infty) \to [0, \infty)$ is convex and nondecreasing. Assume that a quantizer Q with consecutive interval cells $\{S_i\}$ and corresponding codepoints $c_i \in S_i$ is an optimal ECSQ. Then there is a $\lambda \geq 0$ such that for all i, j, if $x \in S_i$, then*

$$\rho(|x-c_i|) - \lambda \log p_i \leq \rho(|x-c_j|) - \lambda \log p_j \tag{4.18}$$

where $p_i = P\{X \in S_i\}$ for all i.

Remark. If $\lambda = 0$, then (4.18) reduces to the nearest neighbor condition, and thus Q is a nearest neighbor quantizer. Moreover, if Q is a nearest neighbor quantizer, then (4.18) holds for any λ if all cells of Q have the same probability, otherwise $\lambda = 0$.

Proof. Denote the subdivision points of Q (that is, the endpoints of the interval cells) by $\{q_i\}$. The conditions of the lemma imply that these points can be ordered such that $q_i < q_{i+1}$ for all i, and without loss of generality we can assume that the cell S_i with corresponding codepoint c_i has the form $S_i = [q_{i-1}, q_i)$. First we show that there is a $\lambda \geq 0$ such that

$$\rho(c_{i+1} - q_i) - \lambda \log p_{i+1} = \rho(q_i - c_i) - \lambda \log p_i \tag{4.19}$$

for all i. Consider the case where Q is a finite-point quantizer. In this case Q is regular, and (4.19) was previously derived by Farvardin and Modestino [7], but without clarifying the conditions on f and ρ. Let Q be an n-point quantizer, let $q_0 = \sigma, q_n = \tau$, and let $\mathbf{q} = (q_1, \ldots, q_{n-1})$ be the subdivision point vector of Q. Since Q is an optimal ECSQ for some constraint $H(Q) \leq R$, it is also optimal among all n-point regular quantizers with codepoints $\{c_1, \ldots, c_n\}$ and entropy less than or equal to R. Denote the distortion and entropy of a quantizer with subdivision point vector $\mathbf{t} = (t_1, \ldots, t_{n-1})$ $(t_i \in (\sigma, \tau))$ and codepoints $\{c_1, \ldots, c_n\}$ by

$$D(\mathbf{t}) = \sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} \rho(|x - c_i|) f(x) dx \quad \text{and} \quad H(\mathbf{t}) = -\sum_{i=1}^{n} p_{\mathbf{t},i} \log p_{\mathbf{t},i}$$

respectively, where $p_{\mathbf{t},i} = P\{t_{i-1} \leq X < t_i\} = \int_{t_{i-1}}^{t_i} f(x)dx$, $t_0 = \sigma$, and $t_n = \tau$. Then $\mathbf{q}$ is a solution of the constrained minimization problem

$$\begin{aligned} &\text{minimize} && D(\mathbf{t}) \\ &\text{subject to} && H(\mathbf{t}) \leq R. \end{aligned}$$

Since both D and H are continuously differentiable functions of $\mathbf{t}$, by the Kuhn-Tucker conditions (see, e.g., [25]) there is a $\lambda \geq 0$ such that the equations

$$\frac{\partial D(\mathbf{t})}{\partial t_i} + \lambda \frac{\partial H(\mathbf{t})}{\partial t_i} = 0, \quad \text{for } i = 1, \ldots, n-1 \tag{4.20}$$

$$\lambda(H(\mathbf{t}) - R) = 0. \tag{4.21}$$

are satisfied for $\mathbf{t} = \mathbf{q}$. The partial derivatives are

$$\begin{aligned} \frac{\partial D(\mathbf{t})}{\partial t_i} &= f(t_i)(\rho(|x - c_i|) - \rho(|x - c_{i+1}|)) \\ \frac{\partial H(\mathbf{t})}{\partial t_i} &= f(t_i)(\log p_{\mathbf{t},i+1} - \log p_{\mathbf{t},i}). \end{aligned}$$

Then, since $f(q_i) > 0$ for $i = 1, \dots, n-1$, by assumption, from (4.20) we obtain (4.19) if Q has a finite number of codepoints.

If p_i is constant for all i, then (4.19) does not depend on λ. Otherwise there is an i such that $p_i \neq p_{i+1}$, and then λ is uniquely given by

$$\lambda = \frac{\rho(c_{i+1} - q_i) - \rho(q_i - c_i)}{\log p_{i+1} - \log p_i}. \tag{4.22}$$

Notice that the numerator in the above expression is independent of the distribution of X, and the denominator depends only on the ratio of the probabilities of consecutive cells. Therefore, if Q is an infinite-point quantizer, then the application of (4.19) to the conditional distributions of X conditioned on $[q_{-n}, q_n)$, $n = 1, 2, \dots$, proves (4.19) for all i (clearly, Q is also optimal for such distributions for its corresponding rate).

Now since ρ is nondecreasing, if $x \in [c_i, q_i]$, then

$$\begin{aligned} \rho(|x - c_i|) - \lambda \log p_i &= \rho(x - c_i) - \lambda \log p_i \\ &\leq \rho(q_i - c_i) - \lambda \log p_i \\ &= \rho(c_{i+1} - q_i) - \lambda \log p_{i+1} \\ &\leq \rho(c_{i+1} - x) - \lambda \log p_{i+1} \end{aligned}$$

where the second equality follows from (4.19). On the other hand, if $x \leq c_i$, then

$$\begin{aligned} \rho(|x - c_i|) - \lambda \log p_i &= \rho(c_i - x) - \lambda \log p_i \\ &\leq \rho(c_i - x) + \rho(q_i - c_i) - \lambda \log p_i \\ &= \rho(c_i - x) + \rho(c_{i+1} - q_i) - \lambda \log p_{i+1} \\ &\leq \rho(c_{i+1} - x) - \lambda \log p_{i+1}. \end{aligned}$$

The second equality holds again by (4.19), and in the last step we used the fact that ρ is convex and nondecreasing. Therefore, if $x \in (q_{i-1}, q_i]$, then

$$\rho(|x - c_i|) - \lambda \log p_i \leq \rho(|x - c_{i+1}|) - \lambda \log p_{i+1}$$

and so for all $j > i$,

$$\rho(|x - c_i|) - \lambda \log p_i \le \rho(|x - c_j|) - \lambda \log p_j.$$

A similar proof shows the corresponding inequality for all $j < i$. $\square$

Combining Corollary 4.4 and Lemma 4.3 with some ideas of [4] allows us to resolve the problem of regularity of optimal ECSQs in the special case of squared error distortion and piecewise monotone and piecewise continuous source densities.

Theorem 4.5 *Let X be a random variable with a density f which is piecewise monotone and piecewise continuous in (σ, τ). Assume that $d(x,y) = (x-y)^2$ and $E\{X^2\} < \infty$. Then for any constraint $R \ge 0$ there is an optimal ECSQ which is regular.*

Proof. Let Q be the optimal ECSQ of Corollary 4.4. If Q is regular, then there is nothing to prove. Assume that Q is not regular. By analyzing the proof of the corollary, one can see that the definition of Q cannot be naturally extended to a point $e \in (\sigma, \tau)$ if and only if Q has two sequences of infinitely many consecutive cells (of positive probability) approaching c from above and from below, respectively. Thus, since Q is not regular, there is a bounded interval I_0 such that Q has infinitely many consecutive cells in I_0. By the definition of piecewise monotonicity and piecewise continuity, I_0 can be partitioned into finitely many intervals such that f is monotone and continuous in each of these intervals. Now Q has infinitely many consecutive cells of positive probability in one of these intervals, say I_1. Therefore, since f is monotone, there is a subinterval $I \subset I_1$ such that f is positive, monotone, and continuous in I, and Q has infinitely many consecutive cells in I. Without loss of generality we can assume that f is nonincreasing on I (the case where f is nondecreasing can be treated similarly). Then by the construction of Q, the cells of Q belonging to I can be assumed to have the form $[q_{i-1}, q_i)$ where $q_{i-1} < q_i$ and $q_i \in I$ for all $i = 1, 2, \ldots$ (also, $q_0 \in I$). Denote the corresponding codepoints by c_i. Then the centroid rule (4.3) implies that $q_{i-1} \le c_i < q_i$, and so by Lemma 4.3, there is a $\lambda \ge 0$ such that for all $i, j \ge 1$ and $x \in [q_{i-1}, q_i)$ we have

$$\rho(|x - c_i|) - \lambda \log p_i \le \rho(|x - c_j|) - \lambda \log p_j$$

where $p_i = P\{X \in [q_{i-1}, q_i)\}$ for all i. If $\lambda > 0$, then, following an idea of [4], for any i we obtain

$$\begin{aligned} -\lambda \log p_i &= \rho(|c_i - c_i|) - \lambda \log p_i \\ &\le \rho(|c_i - c_1|) - \lambda \log p_1 \\ &\le \rho(\ell) - \lambda \log p_1. \end{aligned}$$

where ℓ denotes the length of I. Therefore $-\log p_i$ is bounded from above; hence there is an $\epsilon > 0$ such that $p_i > \epsilon$ for all i. But then $\sum_{i=1}^{\infty} p_i = \infty$, which is impossible. Thus $\lambda = 0$. Then by the remark after Lemma 4.3, the cells $[q_{i-1}, q_i)$ satisfy the nearest neighbor condition. Since $d(x, y) = (x - y)^2$ is the squared error distortion, the optimal codepoint c_i corresponding to $[q_{i-1}, q_i)$ is the centroid of $[q_{i-1}, q_i)$, and since f is nonincreasing, $c_i - q_{i-1} \le q_i - c_i$. Because the cells satisfy the nearest neighbor condition, $c_{i+1} - q_i = q_i - c_i$, and so $q_i - q_{i-1} \le q_{i+1} - q_i$ for all i. That is, the cell lengths increase with i. Therefore, $q_i \to \infty$, contradicting the fact the I is bounded and $q_i \in I$ for all i. Thus Q is regular. $\square$

It is of obvious (practical) interest to determine whether an optimal quantizer has finitely many cells. For the squared error distortion Chou and Betts [4] proved that if the source distribution has lighter tail than that of a Gaussian, then an optimal ECSQ which achieves the lower convex hull of $D_h(R)$ has finitely many cells.

In Proposition 4.1 we attempt (but do not entirely succeed) to extend this result to any optimal ECSQ for a certain class of source distributions.

For a regular quantizer Q we will call the cells $\{S_i\}$ positive (resp. negative) tail cells if $S_1 < S_2 < \cdots$ and $\lim_{n\to\infty} \sup_n S_n = \infty$ (resp. $S_1 > S_2 > \cdots$ and $\lim_{n\to\infty} \inf S_n = -\infty$). If X has a density f which is continuous and positive on the union of the positive (resp. negative) tail cells, then by Lemma 4.3 there is a $\lambda_+ \geq 0$ (resp. $\lambda_- \geq 0$) such that the tail cells satisfy the modified nearest neighbor condition (4.18) with λ_+ (resp. λ_-).

Proposition 4.1 *Let X be a random variable with a density f such that f is finitely piecewise continuous in the interval (σ, τ) and the set $\{x \in (\sigma, \tau) : f(x) = 0\}$ is a union of finitely many intervals and finitely many distinct points. Suppose that $d(x, y) = \rho(|x-y|)$, where $\rho : [0, \infty) \to [0, \infty)$ is a convex and nondecreasing function, and assume that the regular quantizer Q is an optimal ECSQ. Let $\lambda = \min(\lambda_+, \lambda_-) > 0$. Then if for some $\epsilon > 0$*

$$P\{|X| \geq t\} = o(e^{-\rho((1+\epsilon)t)/\lambda}) \tag{4.23}$$

then Q has only a finite number of cells, where $a(t) = o(b(t))$ means $\lim_{t\to\infty} a(t)/b(t) = 0$.

Remark. Specialized to the squared error distortion case, if the source distribution has lighter tail than that of the Gaussian distribution, and if the tail cells of an optimal ECSQ do not satisfy the nearest neighbor condition, then the quantizer has finitely many codepoints. The limitation of this result is that we have been unable to prove that an optimal nearest neighbor ECSQ has only a finite number of codepoints.

Proof. If Q is a finite-point quantizer, then there is nothing to prove. Assume that Q is an infinite-point quantizer. Let $(\ldots, q_{-1}, q_0, q_1, \ldots)$ be the subdivision point vector of Q, and denote the codepoint corresponding to the cell $[q_{i-1}, q_i)$ by c_i. Let $p_i = P\{X \in [q_{i-1}, q_i)\}$.

First assume that f is positive and continuous in (σ, τ). By Lemma 4.3 and the assumptions of the theorem, for all i and $x \in [q_{i-1}, q_i)$ we have

$$\rho(|x - c_i|) - \lambda \log p_i \leq \rho(|x - c_0|) - \lambda \log p_0. \tag{4.24}$$

Let $t_i = 0$ if $q_{i-1} \leq 0 \leq q_i$, and $t_i = \operatorname{sgn}(q_i) \min\{|q_{i-1}|, |q_i|\}$ otherwise, where $\operatorname{sgn}(\cdot)$ is the sign function. Then by (4.24) we have

$$\rho(|t_i - c_i|) - \lambda \log p_i \leq \rho(|t_i - c_0|) - \lambda \log p_0.$$

But $\rho(|t_i - c_i|) > 0$ and $p_i = P\{X \in [q_{i-1}, q_i)\} \leq P\{|X| \geq |t_i|\}$, and hence

$$-\lambda \log P\{|X| \geq |t_i|\} \leq \rho(|t_i - c_0|) - \lambda \log p_0.$$

Equivalently,

$$P\{|X| \geq |t_i|\} \geq p_0 e^{-\rho(|t_i - c_0|)/\lambda}.$$

Now if $|t_i|$ is large enough, then $|(1+\epsilon)t_i| > |t_i - c_0|$, and so

$$P\{|X| \geq |t_i|\} \geq p_0 e^{-\rho((1+\epsilon)t_i)/\lambda}$$

for all sufficiently large $|t_i|$. But (4.23) yields

$$P\{|X| \geq t\} < p_0 e^{-\rho((1+\epsilon)t)/\lambda}$$

if t is large enough, and so there is a $T > 0$ such that $|t_i| < T$ for all i. Then since the cells are pairwise disjoint, there can be at most two cells not contained in $[-T, T]$. Since Q is regular it has only finitely many cells in $[-T, T]$. Thus Q is indeed a finite-point quantizer.

To extend the proof to every density of the theorem, simply apply the same proof for the tail cells of Q, and observe that because of regularity, Q can have only finitely many cells intersecting with the bounded interval between the two tail regions. $\square$

4.4 A discussion of regularity

In this section we summarize the steps from arbitrary quantizers to regular quantizers.

The main tool to deal with this problem is Lemma 4.1. Using the simple procedure described there, we can construct a new quantizer Q for any original quantizer with the same cell probabilities and equal or less distortion (allowing the quantizer not to be defined at some points). Obviously, we can define Q arbitrarily on this set; it will not effect the rate or the distortion of the quantizer, but the goal is to extend the definition of Q while keeping its structural advantages.

If Q is a finite-point quantizer, then the problem is easy to resolve: we can arbitrarily assign the endpoints of the interval cells (the possible points where Q is not defined) to one of the cells whose endpoints they are.

The problem is more complicated if Q has infinitely many cells. Then the cells can be located without any structure in the real line, and it can happen, for example, that Q is not defined on an uncountable set. Assuming that the source has a piecewise monotone and continuous density, we could establish some order among the cells of Q, resulting in the quantizers of Corollary 4.4, which are undefined only on a countable set S which has a finite intersection with any bounded interval. Then the regularity was proved using the modified nearest neighbor condition.

The problem of regularity could be easily resolved, if one could prove directly that an optimal quantizer cannot have infinitely many cells in a finite interval. Unfortunately, we have not been able to find such a proof.

Chapter 5

Conclusions and future work

In this thesis scalar quantization subject to an entropy constrained has been studied.

In Chapter 3 we determined analytically the optimal ECSQ for a source which is uniformly distributed over a bounded interval for a rather wide class of difference distortion measures including rth power distortions $d(x,y) = |x-y|^r$ with $r > 0$. This is the first case when an analytic formula for the optimal ECSQ was rigorously proved. Based on this result we gave a parametric representation of the operational distortion-rate curve $D_h(R)$, and determined some of its properties. This example also showed that the $D_h(R)$ curve is not convex in general, and that an ECSQ achieving the lower convex hull of $D_h(R)$ is rather an exception than the rule.

The structural properties of optimal ECSQs were investigated in Chapter 4. We presented new results concerning the structure of ECSQs. We showed that for continuous sources and convex difference distortion measures any finite-point ECSQ can be "regularized" so that the resulting regular quantizer has the same entropy and equal or less distortion. This result implies that any ECSQ can be arbitrarily well approximated by finite-level regular ECSQs. This result was extended to infinite-point quantizers, using the notion of almost regular quantizers (such quantizers have interval cells and they are defined almost everywhere with respect to the source distribution). We proved that for any quantizer there is an almost regular quantizer with not greater entropy and equal or less distortion.

The problem of the existence of optimal ECSQs was also addressed in Chapter 4. We proved that for convex difference distortion measures and continuous densities an optimal ECSQ exists which is almost regular for any rate constraint $R \geq 0$. Specialized to the case of squared error distortion, we showed that an optimal regular ECSQ exists for all $R \geq 0$ if the source has a piecewise monotone and piecewise continuous density.

In the future it is of interest to extend our results to more general source distributions and distortion measures, and in particular to vector quantizers.

It would be also interesting to clarify under what conditions an optimal ECSQ has finitely many codepoints. An upper bound on the number of codepoints depending on the rate constraint and the source distribution could certainly have significant role in practical quantizer design procedures. Preliminary results concerning the latter problems were provided at the end of Section 4.3.

Appendix

A.1 Polytopal vector quantizers

We call a set $S \subset \mathbb{R}^k$ a *convex polytope*, if it can be obtained as an intersection of finitely many halfspaces in $\mathbb{R}^k$. Then the following lemma can be proved.

Lemma A.1 *If Q is a finite-point quantizer, then its cells $S_1, \ldots, S_N$ are convex sets if and only if they are convex polytopes.*

Proof. Convex polytopes are convex sets, so one direction of the lemma is trivial. To prove the other direction, assume that S_i is a convex subset of $\mathbb{R}^k$ for $i = 1, \ldots, N$. Now for any $i \neq j$, there is a closed hyperplane $H_{i,j}$ such that the interiors of S_i and S_j are on different sides of $H_{i,j}$ (see, e.g., [25]). Let $U_i^{(j)}$ denote the closed halfspace generated by $H_{i,j}$ which contains $\mathrm{int}S_i$, the interior of S_i. Then $\mathrm{int}S_i \subset \bigcap_{j\neq i} U_i^{(j)}$. On the other hand, $\bigcap_{j\neq i} U_i^{(j)} \subset \mathrm{cl}S_i$, where $\mathrm{cl}S_i$ denotes the closure of S_i, since $\mathbb{R}^k \setminus \bigcup_{j\neq i} S_j = S_i$ and $\mathrm{int}\bigcap_{j\neq i} U_i^{(j)} \subset \mathbb{R}^k \setminus \bigcup_{j\neq i} S_j$. (Since the $U_i^{(j)}$ are closed, this implies that $\bigcap_{j\neq i} U_i^{(j)} = \mathrm{cl}S_i$.) Therefore, S_i is a convex polytope. $\square$

A.2 Modified nearest neighbor condition

In this section the proof of the modified nearest neighbor condition for an optimal ECVQ achieving the lower convex hull of $D_h(R)$ is reproduced from [4].

Let Q be an optimal ECVQ achieving the lower convex hull of $D_h(R)$, and denote the cells and corresponding codepoints of Q by $\{S_1, S_2, \dots\}$ and $\{c_1, c_2, \dots\}$, respectively. Let $p_i = P\{X \in S_i\}$. Since Q lies on the convex hull and $D_h(R)$ is nonincreasing, there is a Lagrange multiplier $\lambda \geq 0$ such that Q minimizes $D(Q^*) + \lambda H(Q^*)$ over all possible quantizers Q^* (see also [5]). Note that any λ which is the negative of the slope of a tangent line to the lower convex hull at the point $H(Q)$ is suitable. For example, λ can be chosen to be the negative of the left or right derivative of the lower convex hull at $H(Q)$. Then Q satisfies the *modified nearest neighborhood condition* with parameter λ:

Lemma A.2 *[4] Assume that the quantizer Q achieves the lower convex hull of $D_h(R)$, and denote the negative of the right derivative of the lower convex hull at $H(Q)$ by λ ($\lambda \geq 0$). Then Q minimizes $D(Q) + \lambda H(Q)$ over all quantizers, and so for μ-almost all $x \in \mathbb{R}^k$*

$$d(x, c_{\alpha(x)}) - \lambda \log p_{\alpha(x)} = \min_i d(x, c_i) - \lambda \log p_i$$

where $\alpha(x) = i$ if and only if $Q(x) = c_i$.

Remark. Note that $\min_i d(x, c_i) - \lambda \log p_i$ exists for all $x \in \mathbb{R}^k$, since $\log p_i \to \infty$ if $p_i \to 0$, and so there are always only finitely many indexes i such that $d(x, c_i) - \lambda \log p_i \leq d(x, c_1) - \lambda \log p_1$.

Proof. By the discussion preceding the proof, Q minimizes $D(Q) + \lambda H(Q)$. Assume that on some subset of $\mathbb{R}^k$ with positive probability, $d(x, c_{\alpha(x)}) - \lambda \log p_{\alpha(x)} >$

$\min_i d(x, c_i) - \lambda \log p_i$. Let $\hat{\alpha}(x) = \arg\min_i d(x, c_i) - \lambda \log p_i$, and define $\widehat{Q}$ to have cells $\widehat{S}_i = \{x \in \mathbb{R}^k : \hat{\alpha}(x) = i\}$ and codepoints c_i. Then

$$\begin{aligned} D(Q) + \lambda H(Q) &= E\{d(X, Q(X)) - \lambda \log p_{\alpha(X)}\} \\ &> E\{d(X, \widehat{Q}(X)) - \lambda \log p_{\hat{\alpha}(X)}\} \\ &\geq E\{d(X, \widehat{Q}(X)) - \lambda \log \hat{p}_{\hat{\alpha}(X)}\} \\ &= D(\widehat{Q}) + \lambda H(\widehat{Q}) \end{aligned}$$

where $\hat{p}_i = P\{X \in \widehat{S}_i\}$, and the last inequality follows from the fact that $\sum_i \hat{p}_i \log(\hat{p}_i / p_i) \geq 0$. Hence Q does not minimize $D(Q) + \lambda H(Q)$, a contradiction. $\square$

A.3 Limits and integrals

In this section we state some well-known results of real analysis used in this thesis corresponding to the interchangeability of limits and integrals of extended real valued functions on $\mathbb{R}$. The proof of the statements can be found in any introductory book on real analysis or measure theory, such as [2] or [29].

Theorem A.1 *(Fatou's Lemma.) Let μ be a measure on $\mathbb{R}$. If $f_1, f_2, \ldots$ are non-negative extended real valued Borel measurable functions, then*

$$\int_{\mathbb{R}} \liminf_{n\to\infty} f_n(x) d\mu(x) \leq \liminf_{n\to\infty} \int_{\mathbb{R}} f_n(x) d\mu(x).$$

Theorem A.2 *(Dominated Convergence Theorem.) Let μ be a measure on $\mathbb{R}$. If $f_1, f_2, \ldots, f, g$ are extended real valued Borel measurable functions such that $|f_n(x)| \leq g(x)$ and $\lim_n f_n(x) = f(x)$ for μ-almost all $x \in \mathbb{R}$, and $\int_{\mathbb{R}} g(x) d\mu(x)$ is finite, then*

$$\lim_{n\to\infty} \int_{\mathbb{R}} f_n(x) d\mu(x) = \int_{\mathbb{R}} f(x) d\mu(x).$$

Theorem A.3 *(Monotone Convergence Theorem.) Let μ be a measure on $\mathbb{R}$. If $f_1, f_2, \ldots$ is a nondecreasing sequence of nonnegative extended real valued Borel measurable functions and $\lim_{n\to\infty} f_n(x) = f(x)$ for μ-almost all $x \in \mathbb{R}$, then*

$$\lim_{n\to\infty} \int_{\mathbb{R}} f_n(x) d\mu(x) = \int_{\mathbb{R}} f(x) d\mu(x).$$

A.4 Proofs of some technical results

This section contains some technical proofs of previous statements of the thesis.

Lemma A.3 *Assume that the distortion measure is of the form $d(x, y) = \rho(|x - y|)$, where $\rho : [0, \infty) \to [0, \infty)$ is a nondecreasing and lower semicontinuous function satisfying (4.2). Then for any $S \subset \mathbb{R}$ there is a $c^* \in \mathbb{R}$ such that*

$$\int_S d(x, c^*) d\mu(x) = \inf_{c\in\mathbb{R}} \int_S d(x, c) d\mu(x) < \infty. \tag{A.1}$$

Proof. First note that (4.2) implies that the infimum in (A.1) is finite. If $\mu(S) = 0$, then any choice of c^* results in zero distortion, and so there is nothing to prove. Assume that $\mu(S) > 0$. Let $\lim_{x\to\infty} \rho(x) = a$. Consider a sequence of real numbers $\{x_n\}$ such that $|x_n| \to \infty$ as $n \to \infty$. Then by Fatou's lemma,

$$\liminf_{n\to\infty} \int_S \rho(|x - x_n|) d\mu(x) \geq \int_S \liminf_{n\to\infty} \rho(|x - x_n|) d\mu(x) = a\mu(S). \tag{A.2}$$

If a is finite, then this shows that for any $\epsilon > 0$ there is a $y > 0$ such that

$$\int_S \rho(|x - c|) d\mu(x) > a\mu(S) - \epsilon \qquad \text{if } |c| > y. \tag{A.3}$$

If $a = \infty$, then it follows from (A.2) that for any $K > 0$ there is a $y > 0$ such that

$$\int_S \rho(|x - c|) d\mu(x) > K \qquad \text{if } |c| > y. \tag{A.4}$$

If

$$\inf_{c\in\mathbb{R}} \int_S \rho(|x-c|)d\mu(x) = a\mu(S)$$

then any choice of c^* suffices. Otherwise (A.3) and (A.4) imply that

$$\inf_{c\in\mathbb{R}} \int_S \rho(|x-c|)d\mu(x) = \inf_{c\in[-y,y]} \int_S \rho(|x-c|)d\mu(x).$$

Thus we only need to show that

$$g(c) = \int_S \rho(|x-c|)d\mu(x)$$

is lower semicontinuous, since in this case $g(c)$ will take its infimum in the compact set $[-y, y]$. But Fatou's lemma and the lower semicontinuity of ρ imply that

$$\liminf_{c_n\to c} \int_S \rho(|x-c_n|)d\mu(x) \geq \int_S \liminf_{c_n\to c} \rho(|x-c_n|)d\mu(x) \geq \int_S \rho(|x-c|)d\mu(x)$$

and so $g(c)$ is lower semicontinuous. $\square$

Proof of (4.11). The codepoint of Q_n corresponding to the cell $\bigcup_{i\geq n} S_{\pi(i)}$ is 0, and hence the distortion of the quantizer Q_n can be written in the form

$$\begin{aligned} D(Q_n) &= \int_{(\sigma,\tau)} \rho(|x-Q_n(x)|)d\mu(x) \\ &= \int_{C_n} \rho(|x-Q_n(x)|)d\mu(x) + \int_{(\sigma,\tau)\setminus C_n} \rho(|x|)d\mu(x). \end{aligned} \tag{A.5}$$

Since the sequence of sets $\{C_n\}$ increases to (σ,τ) and $Q_n(x) = Q(x)$ if $x \in C_n$, we have $\lim_{n\to\infty} \mathbb{I}_{C_n}(x)\rho(|x-Q_n(x)|) = \rho(|x-Q(x)|)$, where $\mathbb{I}_{C_n}(\cdot)$ denotes the characteristic function of the set C_n. Moreover, since $Q(x) \geq 0$, the sequence $\{\mathbb{I}_{C_n}(x)\rho(|x-Q_n(x)|)\}_{n=1}^{\infty}$ is nondecreasing in n for any $x \in \mathbb{R}$. Thus the monotone convergence theorem (see Theorem A.3) implies

$$\lim_{n\to\infty} \int_{C_n} \rho(|x-Q_n(x)|)d\mu(x) = \int_{(\sigma,\tau)} \rho(|x-Q(x)|)d\mu(x) = D(Q). \tag{A.6}$$

On the other hand, $\nu(A) = \int_A \rho(|x|)d\mu(x)$ for all Borel set $A \subset \mathbb{R}$ is a finite measure, since $\nu(\mathbb{R}) = E\{\rho(|X|)\} < \infty$. Thus by the continuity of finite measures,

$$\lim_{n\to\infty} \int_{(\sigma,\tau)\setminus C_n} \rho(|x|)d\mu(x) = 0 \tag{A.7}$$

because $\lim_{n\to\infty}(\sigma,\tau) \setminus C_n = \emptyset$. Now (A.5)–(A.7) imply (4.11). □

A.5 The Cantor ternary set

In this section the construction and some properties of the Cantor ternary set are given (see, e.g., [2]).

Let E_1 be the middle third of the interval $[0, 1]$, that is, $E_1 = (\frac{1}{3}, \frac{2}{3})$. Let E_2 be the union of the middle thirds of the two intervals that remain after E_1 is removed, that is, $E_2 = (\frac{1}{9}, \frac{2}{9}) \cup (\frac{7}{9}, \frac{8}{9})$. Continue the construction; let E_n be the union of the middle thirds of the intervals that remain after $E_1, \ldots, E_{n-1}$ are removed. The Cantor ternary set C is then defined as

$$C = [0, 1] \setminus \bigcup_{n=1}^{\infty} E_n.$$

It is easy to see that $x \in C$ if and only if $x \in [0, 1]$ and x can be expressed in ternary form using only digits 0 and 2. Then it follows that C is *uncountable*. Moreover, notice that the sum of the lengths of the intervals removed is 1. Therefore, C has *Lebesgue measure* 0.

Other interesting properties also follow from the definition: C is closed, perfect (every point of C is a limit point of C), and nowhere dense.

Bibliography

[1] E. A. Abaya and G. L. Wise, "Convergence of vector quantizers with applications to optimal quantization," *SIAM Journal on Applied Mathematics*, vol. 44, pp. 183–189, 1984.

[2] R. B. Ash, *Real Analysis and Probability.* New York: Academic Press, 1972.

[3] T. Berger, "Optimum quantizers and permutation codes," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 759–765, Nov. 1972.

[4] P. A. Chou and B. J. Betts, "When optimal entropy-constrained quantizers have only a finite number of codewords," in *Proc. IEEE Int. Symp. Information Theory*, (Cambridge, MA, USA), p. 97, Aug. 1998.

[5] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-37, pp. 31–42, Jan. 1989.

[6] T. Cover and J. A. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

[7] N. Farvardin and J. W. Modestino, "Optimum quantizer performance for a class of non-gaussian memoryless sources," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 485–497, May. 1984.

[8] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[9] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 373–380, July 1979.

[10] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 676–683, Sep. 1968.

[11] R. M. Gray, *Source Coding Theory.* Boston: Kluwer, 1990.

[12] R. M. Gray, J. C. Kieffer, and Y. Linde, "Locally optimum block quantizer design," *Information and Control*, vol. 45, pp. 178–198, 1980.

[13] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inform. Theory*, vol. IT-44, pp. 2325–2383, Oct. 1998.

[14] A. György and T. Linder, "Optimal entropy constrained scalar quantization of a uniform source," *IEEE Trans. Inform. Theory*, vol. IT-46, pp. 2704–2711, Nov. 2000.

[15] A. György and T. Linder, "Optimal entropy constrained scalar quantization of a uniform source," in *Proc. Int. Symp. Inform. Theory Appl.*, (Honolulu, HI, USA), pp. 85–88, Nov. 2000.

[16] G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities.* Cambridge University Press, 1959.

[17] F. Jones, *Lebesgue Integration on Euclidean Space.* London: Jones and Bartlett, 1993.

[18] J. C. Kieffer, "Exponential rate of convergence for Lloyd's method I," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 205–210, Mar. 1982.

[19] J. C. Kieffer, "Uniqueness of locally optimal quantizer for log-concave density and convex error weighting function," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 42–47, Jan. 1983.

[20] J. C. Kieffer, "Survey of the theory of source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-39, pp. 1473–1491, Sep. 1993.

[21] J. C. Kieffer, T. M. Jahns, and V. A. Obuljen, "New results on optimal entropy-constrained quantization," *IEEE Trans. Inform. Theory*, vol. IT-34, pp. 1250–1258, Sep. 1988.

[22] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. COM-28, pp. 84–95, Jan. 1980.

[23] T. Linder and R. Zamir, "Causal source coding of stationary sources with high resolution," submitted to *2001 IEEE Int. Symp. Inform. Theory.*

[24] S. P. Lloyd, "Least squared quantization in PCM," unpublished memorandum, Bell Lab., 1957; reprinted in *IEEE Trans. Inform. Theory,* vol. IT-28, pp. 129-137, Mar. 1982.

[25] D. G. Luenberger, *Linear and Nonlinear Programming.* Reading, Mass.: Addison-Wesley, 2nd ed., 1984.

[26] J. Max, "Quantizing for minimum distortion," *IEEE Trans. Inform. Theory*, vol. IT-6, pp. 7–12, Mar. 1960.

[27] A. N. Netravali and R. Saigal, "Optimum quantizer design using a fixed-point algorithm," *Bell Syst. Tech. J*, vol. 55, pp. 1423–1435, Nov 1976.

[28] D. L. Neuhoff and R. K. Gilbert, "Causal source codes," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 701–713, Sep. 1982.

[29] O. A. Nielsen, *An Introduction to Integration and Measure Theory.* New York: Wiley, 1997.

[30] D. Pollard, "Quantization and the method of k-means," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 199–205, Mar. 1982.

[31] W. Rudin, *Real and Complex Analysis.* New York: McGraw-Hill, 3rd ed., 1987.

[32] M. J. Sabin, *Global convergence and empirical consistency of the generalized Lloyd algorithm.* PhD thesis, Stanford Univ., 1984.

[33] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.

[34] F. Topsøe, "Instances of exact prediction and a new type of inequalities obtained by anchoring," in *IEEE Information Theory Workshop*, (Kruger National Park, South Africa), p. 99, June 1999.

[35] A. V. Trushkin, "Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting function," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 187–198, Mar. 1982.

[36] A. D. Wyner, "An upper bound on entropy series," *Information and Control*, vol. 20, pp. 176–181, 1972.

[37] J. Ziv, "On universal quantization," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 344–347, May 1985.

Vita

András György

EDUCATION

1999-2000 Queen's University, Kingston, Canada, M.Sc. (Eng.) in Mathematics and Engineering, expected date of graduation: December 2000
Title of thesis: On Optimal Entropy-Constrained Scalar Quantization
Supervisor: Dr. Tamás Linder

1994-1999 Technical University of Budapest, Hungary, M.Sc. in Technical Informatics, with highest honour
Title of thesis: An Analysis of Measurement Based Call Admission Control Algorithms for ATM Networks
Supervisor: Dr. László Györfi

EXPERIENCE

1999-2000 Research Assistant, Queen's University.

1999-2000 Teaching Assistant, Queen's University.

1996-1999 Research Assistant, Technical University of Budapest.

1998 Visiting Researcher, Center for Wireless Communications, University of California, San Diego.

1997 Research Assistant, Analogical and Neural Computing Laboratory of the Computer and Automation Institute of the Hungarian Academy of Sciences (MTA SzTAKI).

1994-1997 Part-time mathematics instructor, Árpád High School, Budapest, Hungary.

PUBLICATIONS

Journal papers:

- A. György and T. Linder, "Optimal Entropy-Constrained Scalar Quantization of a Uniform Source", to appear in *IEEE Transactions on Information Theory*, November 2000.

- A. György, T. Linder, and K. Zeger, "On the Rate-Distortion Function of Random Vectors and Stationary Sources with Mixed Distributions", *IEEE Transactions on Information Theory*, vol. IT-45, pp. 2110–2115, September 1999.

Refereed conference papers:

- A. György and T. Linder, "Optimal Entropy-Constrained Scalar Quantization of a Uniform Source", *2000 International Symposium on Information Theory and Its Applications*, Honolulu, HI, USA, November 2000.

- A. György, T. Linder, and K. Zeger, "On the Rate-Distortion Function of Random Vectors and Stationary Sources with Mixed Distributions", *6th Canadian Workshop on Information Theory*, Kingston, Ont, Canada, June 1999.

- A. György, T. Linder, and K. Zeger, "Lossy Coding of Sources with Mixed Distribution," *33rd Conference on Information Sciences and Systems*, Baltimore, MD, USA, March 1999.

- A. György, "Upper Bounds on the Tail Distribution of a Single Server Queue", *High Speed Networking'98 Winter Workshop*, Budapest, Hungary, December 1998.

Non-refereed conference papers:

- A. György, "Measurement Based Call Admission Control Algorithms for ATM Networks", *V. Young Statisticians Meeting*, Udine, Italy, October 2000.

- A. György, "Upper Bounds on the Tail Distribution of a Single Server Queue", *Scientific Conference for Students at the Faculty of Electrical Engineering and Informatics, Technical University of Budapest*, Budapest, Hungary, November 1998.

- A. György, "On the Theoretical Limits of Lossy Source Coding", *Scientific Conference for Students at the Faculty of Electrical Engineering and Informatics, Technical University of Budapest*, Budapest, Hungary, November 1998.

- A. György, "Upper Bounds on the Cell Loss Probability in ATM Networks" (in Hungarian), *Scientific Conference for Students at the Faculty of Electrical Engineering and Informatics, Technical University of Budapest*, Budapest, Hungary, November 1996.

AWARDS

2000 First prize winner at the Master's Thesis Competition of the Scientific Association for Infocommunications of Hungary.

1999 Queen's Graduate Award.

1999 Third prize winner at the Excellent Master's Thesis Competition of the IMFA Foundation.

1998 First prize winner (twice) at the Scientific Conference of Students at the Faculty of Electrical Engineering and Informatics, Technical University of Budapest.

1997 Award of the Professional Resource International.

1996 First prize winner at the Scientific Conference of Students at the Faculty of Electrical Engineering and Informatics, Technical University of Budapest.

SCHOLARSHIPS

2000 Ontario Graduate Scholarship.

1999 Scholarship of the Bizáki Puki Péter Memorial Foundation.

1999 Scholarship of the Soros Foundation.

1997-1999 Scholarship of the Hungarian Republic.

1996 Scholarship of the Faculty of Electrical Engineering and Informatics, Technical University of Budapest.